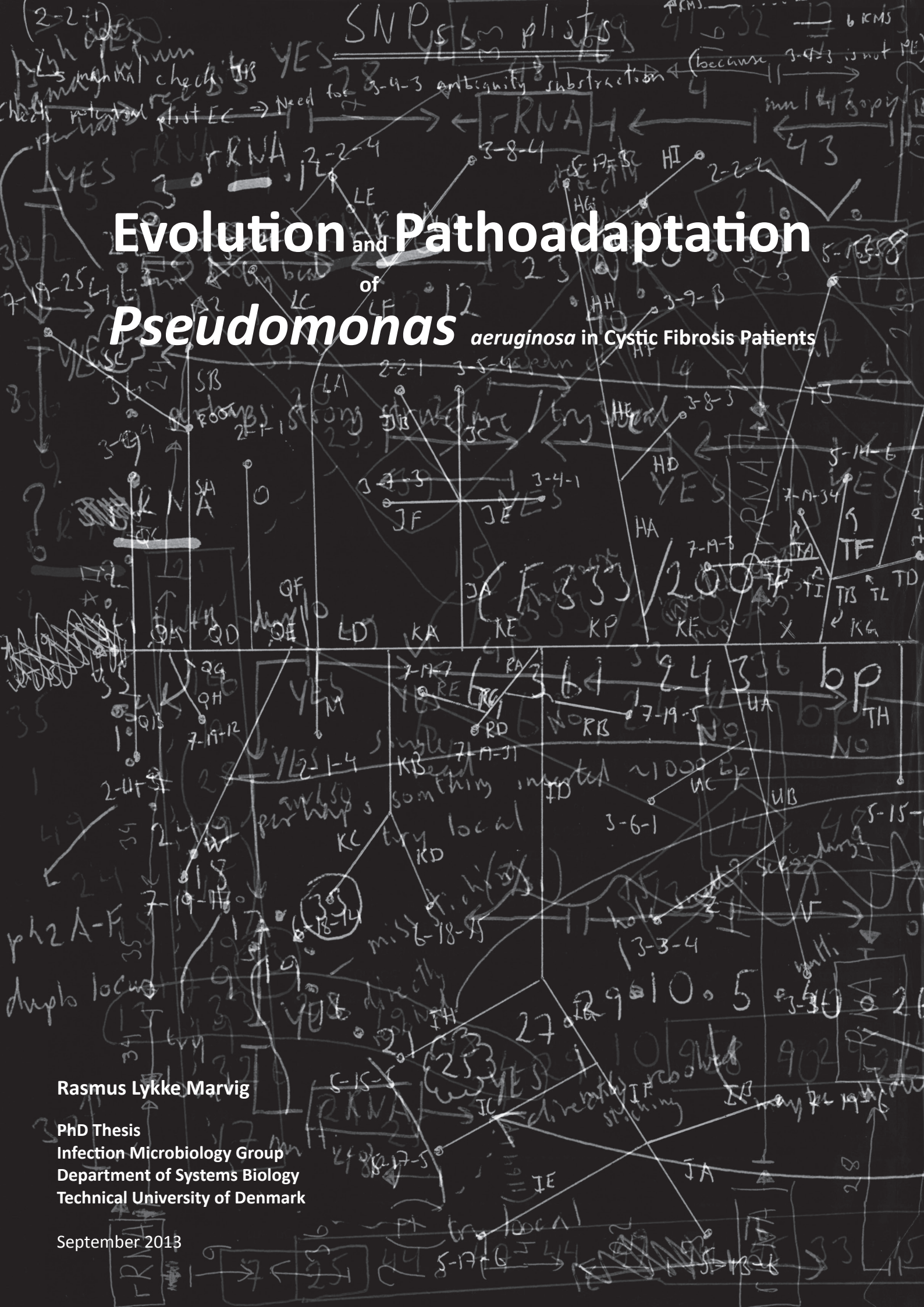
Evolution and Pathoadaptation of Pseudomonas aeruginosa in Cystic Fibrosis Patients

Marvig, Rasmus Lykke

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

Evolution and Pathoadaptation of *Pseudomonas aeruginosa* in Cystic Fibrosis Patients



Rasmus Lykke Marvig

PhD Thesis
Infection Microbiology Group
Department of Systems Biology
Technical University of Denmark

September 2013

Evolution and Pathoadaptation of
Pseudomonas aeruginosa in Cystic Fibrosis Patients

ISBN: 978-87-91494-58-1
Infection Microbiology Group
Department of Systems Biology
Technical University of Denmark

Cover art by [Noerkjaer] “Collage of original project sketches”

“Life is beautiful“

Preface

This thesis is written as a partial fulfillment of the requirements to obtain a PhD degree at the Technical University of Denmark. The work presented in the thesis was carried out from October 2010 to September 2013 at the Infection Microbiology Group (IMG), Department of Systems Biology, Technical University of Denmark under the supervision of Professor Søren Molin (supervisor) and Associate Professor Lars Jelsbak (co-supervisor). The work was funded by a PhD stipend from the Technical University of Denmark.

Rasmus Lykke Marvig

Kongens Lyngby, September 2013

Abstract

Molecular and mechanistic understanding of evolution is essential for our ability to comprehend the development of life on Earth. Life appeared around 4 billion years ago, and has ever since adapted and diversified through the process of evolution. The focus of this thesis has been to increase our understanding of how bacteria evolve and genetically adapt in a natural environment. In particular we sought to identify the genes that are targeted by mutation to optimize fitness in a given environment, and to understand the evolutionary mechanisms that govern the genetic change.

Pseudomonas aeruginosa is the dominating pathogen of chronic airway infections in patients with cystic fibrosis (CF), and the bacterial long-term persistence in CF hosts involves mutation and selection of genetic variants with increased fitness in the CF airways. We performed a retrospective study of the *P. aeruginosa* DK2 clone type, which is a transmissible clone isolated from chronically infected Danish CF patients over a period of 38 years. Whole-genome analysis of DK2 isolates enabled a fine-grained reconstruction of the recent evolutionary history of the DK2 lineage and an identification of bacterial genes targeted by mutations to optimize pathogen fitness. The identification of such pathoadaptive genes gives new insight into how the pathogen evolves under the selective pressures of the host immune system and drug therapies. Furthermore, isolates with increased rates of mutation (hypermutator phenotype) emerged in the DK lineage. While this phenotype may accelerate evolution, we also showed that hypermutators display differential mutagenesis of certain genes which enable them to follow alternative evolutionary pathways.

Overall, our study identifies genes important for bacterial adaptation to a human host environment and provides insight into the different mutational mechanisms that govern the adaptive genetic changes.

Dansk resumé

Molekylær og mekanistisk forståelse af evolution er essentiel for vores evne til at begribe livets udvikling på Jorden. Livet opstod for omkring 4 milliarder år siden og har lige siden tilpasset og forskelliggjort sig gennem den evolutionære proces. Fokus i denne afhandling har været at øge vores forståelse af, hvordan bakterier udvikler og genetisk tilpasser sig i naturlige omgivelser. Vi søgte især efter at identificere gener, som er mål for mutationer, der optimerer fitness I et givet miljø, og at forstå de evolutionære mekanismer som styrer den genetiske forandring.

Pseudomonas aeruginosa er den dominerende patogen i kroniske luftvejsinfektioner hos patienter med cystisk fibrose (CF), og langstrakt bakteriel vedbliven i CF værter involverer mutationer og selektion af genetiske varianter med øget fitness i CF luftvejene. Vi udførte et tilbageskuende studie af *P. aeruginosa* DK2 klontypen, som er en overførbar klontype, der er isoleret fra kronisk inficerede danske CF patienter over periode på 38 år. Analyse af den samlede arvemasse hos DK2 isolaterne muliggjorde en finkornet rekonstruktion af DK2 linjens nylige evolutionære historie og identifikation af bakteriegener, som var mål for mutationer, der optimerede patogen-fitness. Identifikationen af sådan pathoadaptive gener giver ny indsigt i, hvordan patogenet udvikler sig under selektionspresset fra værten og lægemiddelsbehandling. Endvidere opstod der isolater med forøgede mutationshastigheder (hypermutator fænotype) i DK linjen. Mens denne fænotype måske kan accelerere evolutionen viste vi også, at hypermutatorer udviser uens mutagenese af særlige gener, hvilket gør dem i stand til at følge alternative evolutionære veje.

Samlet set identificerede vores studie gener, som er vigtige for bakteriel tilpasning til et menneskeligt værtsmiljø, og gav indsigt i de forskellige mutationsmekanismer som styrer de tilpassende genetiske forandringer.

Contents

Acknowledgements

The era of genomics has gained pace with the advent of next generation sequencing (NGS) technologies. I would like to thank Professor Søren Molin for letting me into the front seat of IMG's attempt to take advantage of these technologies. Søren is truly visionary: when he in 2010 asked me to sequence and analysis the first 12 genomes of *P. aeruginosa*, I honestly thought that he was "nuts". Nonetheless, I was curious to give it a try in collaboration with Morten Sommer at Harvard University (who do not want to go there?), and today we have continued to use NGS for genomics and transcriptomics on close to a thousand bacterial samples.

All the science that has come out of our efforts is highly collaborative. I would like to thank Helle Johansen and Niels Høiby for providing the unique collection of clinical isolates of *P. aeruginosa*, which has served as a basis for my investigations, and Lars Jelsbak and Lei Yang for the initial work on this collection. Most focus has been on the DK2 lineage, and I would like to thank Martin Rau for his help describing the genomic content of this lineage. In parallel to my project, Trine Markussen has driven the investigation of the DK1 lineage, and many of our findings have been promoted by reflection on her results.

Thanks to Søren Damkiær for his generous contribution of ideas, illustrations, humor, computer know-how, and crucial expertise in the "wet"-lab. Søren and I co-supervised two master students: Mette Søndergaard and Hossein Khademi. Both of them are pivotal authors on the resulting research articles, thanks.

I would like to express my gratitude to the collaborators on the many other projects that I have been involved in during my PhD. This includes María Gómez-Lozano and Katherine Long, Nicholas Jochumsen and Anders Folkesson, and Sofía Feliziani who helped me to learn about transcriptomics, laboratory evolution experiments, and hypermutators, respectively. Also, I am grateful to Professor Edward Delong who accepted me to visit his group at Massachusetts Institute of Technology to learn about marine microbiology and metatranscriptomics. This 5-months research stay was made possible through a Ministry of Science's EliteForsk Travel Scholarship. Furthermore, I would like acknowledge Lea Madsen and Helle Johansen for their efforts which

have made it possible for us to expand our collection of genome sequenced isolates to cover >50 clone types of *P. aeruginosa*.

Beyond the scientific contributions, I would like to thank Thomas Rasmussen for teaching me that it is ok to strive for the top. Nonetheless, as a consequence my friends and family often have to listen to a megalomaniac testing his ideas on the safe home field. In this regard, especially thanks to my office roomie Vinoth Vigneswaran and my beloved Maria, who also took the journey to Boston with me.

Finally, thanks to my supervisors Søren and Lars for their attentive support. They have not only served as excellent teachers, but also they have appeared as role models representing a set of beliefs that I find most valuable in my future work as a scientist.

Research articles that are included in this thesis

Yang L*, Jelsbak L*, **Marvig RL**, Damkiær S, Workman CT, Rau MH, Hansen SK, Folkesson A, Johansen HK, Ciofu O, Høiby N, Sommer MO, Molin S. Evolutionary dynamics of bacteria in a human host environment. *Proc Natl Acad Sci U S A*. 2011 May 3;108(18):7481-6.

Marvig RL*, Søndergaard MS*, Damkiær S, Høiby N, Johansen HK, Molin S, Jelsbak L. Mutations in 23S rRNA confer resistance against azithromycin in *Pseudomonas aeruginosa*. *Antimicrob Agents Chemother*. 2012. 56(8):4519-21.

Rau MH*, **Marvig RL***, Ehrlich GD, Molin S, Jelsbak L. Deletion and acquisition of genomic content during early stage adaptation of *Pseudomonas aeruginosa* to a human host environment. *Environ Microbiol*. 2012 Aug;14(8):2200-11.

Marvig RL, Johansen HK, Molin S, Jelsbak L. Genome Analysis of a Transmissible Lineage of *Pseudomonas aeruginosa* Reveals Pathoadaptive Mutations and Distinct Evolutionary Paths of Hypermutators. *PLoS Genet*. 2013. 9(9):e1003741.

Marvig RL*, Damkiær S*, Khademi H, Markussen, Molin S, Jelsbak L. Within-Host Evolution of Pseudomonas aeruginosa Reveals Adaptation Towards Iron Acquisition from Hemoglobin. *Unpublished*. 2013.

* Denotes equal contribution.

Published works that are not included in this thesis

Marvig RL, Jochumsen N, Johansen HK, Høiby N, Molin S, Sommer MO, Jelsbak L, Folkesson A. Draft genome sequences of *Pseudomonas aeruginosa* B3 strains sampled from cystic fibrosis patient undergoing antibiotic chemotherapy. Accepted for publication in *Genome Announce*. 2013.

Gómez-Lozano M, **Marvig RL**, Molin S, Long KS. Identification of bacterial small RNAs by RNA sequencing. *Methods in Pseudomonas aeruginosa*. In press; 2013 Apr pp. 1–34.

Gómez-Lozano M, **Marvig RL**, Molin S, Long KS. Genome-wide identification of novel small RNAs in *Pseudomonas aeruginosa*. *Environ Microbiol*. 2012. 14(8):2006-16.

Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, **Marvig RL**, Jelsbak L, Sicheritz-Pontén T, Ussery DW, Aarestrup FM, Lund O. Multilocus sequence typing of total-genome-sequenced bacteria. *J Clin Microbiol*. 2012. 50(4):1355-61.

Marvig RL, Blokesch M. Natural transformation of *Vibrio cholerae* as a tool - optimizing the procedure. *BMC Microbiol*. 2010. 10:155.

Abbreviations

CF	Cystic fibrosis
CFTR	Cystic fibrosis transmembrane conductance regulator
DNA	Deoxyribonucleic acid
dN	Rate of nonsynonymous genetic change
dS	Rate of synonymous genetic change
HGT	Horizontal gene transfer
IMG	Infection microbiology group
IS	Insertion sequence
kpb	Kilo base pair
LPS	Lipopolysaccharide
Mbp	Mega base pair
MP	Maximum-parsimony
NGS	Next generation sequencing
PFGE	Pulsed field gel electrophoresis
phu	*Pseudomonas* heme utilization
ROS	Reactive oxygen species
SNP	Single nucleotide polymorphism
WGS	Whole-genome sequencing

List of figures

Chapter 1

Introduction and thesis overview

A molecular and mechanistic understanding of evolution is central for our ability to comprehend the development of life on Earth. Life appeared around 4 billion years ago [1], and since then life has adapted and diversified through the process of evolution to withstand environmental changes and to spread to new niches.

Bacteria constitutes the most abundant and diverse form of life on Earth, and these unicellular organisms show a remarkably ability to evolve and genetically adapt to changing environments. The diversity and adaptability of bacteria offer a great potential in industrial applications, but the same features make the bacteria a threat as infectious agents. In both industrial and clinical settings, however, insight into the evolution and genetic adaptation of bacteria is important for our ability to engineer bacteria and fight bacterial infections.

The focus of this thesis has been to increase our understanding of how bacteria evolve and genetically adapt in a natural environment. In particular we sought: (a) to identify the genes that are targeted by mutation to optimize fitness in a given environment; and (b) to understand the evolutionary mechanisms that govern the genetic change.

The model system used for our investigation has been long-term chronic airway infections in cystic fibrosis (CF) patients caused by the opportunistic pathogen *Pseudomonas aeruginosa*. In this model system we are able to follow the molecular evolution of *P. aeruginosa* within its human host, and thereby decipher the genetic adaptation of *P. aeruginosa* to the human host after transition from its environmental habitat.

1.1 Thesis outline

This thesis is organized into six chapters. While the current chapter (**Chapter 1**) introduces the thesis, **Chapter 2** gives a general introduction to molecular evolution. This includes an introduction to mutation, natural selection, and genetic drift that all

together serve to explain the principles and mechanisms of genetic adaptation. In addition, **Chapter 2** also describes the phylogenetic methods that have been used in the research articles presented later in the thesis. **Chapter 3** presents how whole-genome sequencing (WGS) of clinical collections of bacteria offers an opportunity to study microevolution, *i.e.* evolutionary events between closely related organisms. This leads to an introduction of *P. aeruginosa* infections in CF patients that represent the infectious-disease model, which we use to study molecular evolution and genetic adaptation *in vivo*. A special emphasis has been put on the description of pathogen phenotypes and host factors relevant to the findings of this thesis. The thesis' findings are then presented in **Chapter 4**. This includes an introduction to the background leading to the thesis aims and the context of the resulting research articles. The research articles themselves are enclosed in full length in **Chapter 6**, and **Chapter 5** provides an overall conclusion and perspective of the thesis.

Chapter 2

Molecular evolution

Evolution is the process by which life has evolved from a single origin to the vast diversity of organisms present today. To understand life as it is today, it is therefore essential to study the evolutionary processes that have shaped it.

All living organisms carry hereditary genetic material stored in the form of deoxyribonucleic acid (DNA), which they pass on to their progeny. The entirety of an organism's hereditary material is referred to as the genome. The genome encodes the genetic instructions used for the functioning of the organism, and as such the underlying genome ultimately determines the phenotypic repertoire of an organism. Thereby, understanding molecular evolution is a keystone in understanding the evolutionary process itself.

2.1 Mutation, natural selection, and genetic drift

The genetic content of organisms can change either by *de novo* mutation of already present DNA or by acquisition of DNA from other organisms, *i.e.* horizontal gene transfer (HGT). Mutation of present DNA includes single nucleotide substitutions, insertions, deletions, and rearrangements such as duplications, inversions, and translocations. Mutations are stochastic, and they result from errors in DNA replication and repair, recombination, exposure to mutagens, or mobile genetic elements.

Mutations can either be beneficial, neutral, or detrimental to the reproductive success, *i.e.* the fitness, of an organism (or rather a population of organisms) in its environment. Natural selection will act on the genetic diversity created by mutation within a population to increase the frequency of beneficial genetic variants till they eventually become fixed in the population. Nonetheless, neutral and even detrimental mutations can become fixed in a population due to genetic drift or hitchhiking.

Genetic drift is the change in the frequency of a genetic variant due to random sampling of offspring from a population (**Figure 1**), whereas hitchhiking refers to an increase of a genetic variant by virtue of being genetically linked to a beneficial genetic variant in another locus. Hitchhiking on beneficial mutations is particular strong in asexual populations in which the entire genome acts as a single a linkage-group.

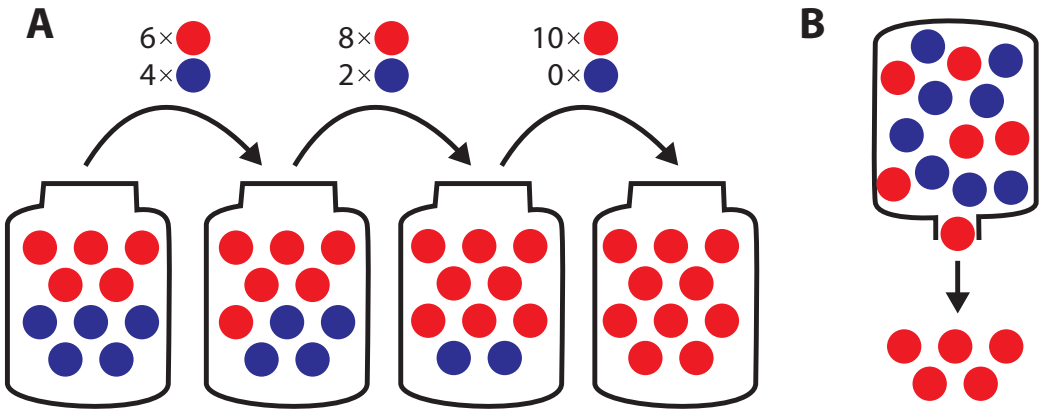


Figure 1: Genetic drift and bottle-neck effects. (A) The principle of genetic drift can be illustrated by sampling of marbles from a jar where marbles with different colors represents different genetic variants. If sampling of "offspring" marbles is done randomly and with replacement, a random shift in the frequency of the different colors might occur in the resulting jar of marbles. Several rounds of such random sampling can therefore result in a jar of marbles with only a single color and with the other color being lost permanently. The effect of genetic drift is dependent on the population size. Thus, the effect of genetic drift is large for small population sizes and negligible for very large population sizes. (B) Nonetheless, even large populations can encounter a sudden decrease in population size, *i.e.* a bottle-neck, leading to radical changes in allele frequencies by genetic drift.

Identification of genetic variation is not *per se* followed by information on if the variation has become fixated due to natural selection or genetic drift. Nonetheless, general conclusions about to what extent natural selection has been the driving force in the fixation of genetic variants, can be inferred from measuring the relative rates of nonsynonymous (dN) and synonymous (dS) genetic changes, where nonsynonymous changes alter the amino acid sequence of encoded proteins and synonymous changes do not. This measure is calculated based on substitution mutations, and its logic is that

only nonsynonymous substitutions (missense and nonsense substitutions) are potential subjects to natural selection, whereas synonymous substitutions (silent substitutions) are regarded as neutral and therefore not subjects of natural selection. The dN/dS ratio can be interpreted as the selection pressure acting on the protein coding genome [2]. A ratio greater than one implies that there has been a positive selection for mutations; whereas a ratio less than one implies that there has been a selection for removal of mutations, *i.e.* negative selection. However, note that the dN/dS ratio should be interpreted as an average signal of selection not giving any information on if both positive and negative selection has acted at different times during the evolution.

$$\frac{dN}{dS} = \frac{\text{No. of fixated nonsynonymous substitutions}}{\text{No. of fixated synonymous substitutions}} \times \frac{\text{No. of synonymous sites}}{\text{No. nonsynonymous sites}}$$

Equation 1: Equation used to calculate the rate of nonsynonymous substitutions relative to the rate of synonymous substitutions.

In this thesis the dN/dS ratio is calculated based on the simplistic assumption that the rates of different base substitutions are the same. Given this, the codon usage in *P. aeruginosa* predicts a 25% probability that a random mutation would be synonymous, and we can calculate the dN/dS according to Equation 1. However, note that this is an approximate measure, as it does not take into account selection on codon usage, mutational skews, selection for higher GC content, and bottle-neck effects [2, 3]. Nonetheless, for the conclusions of this thesis we find it most transparent and appropriate to use a simple and approximate measure. At the same time, we compare our findings to Smith *et. al.* (2006) and Barrick *et al.* (2009) who also assume a random distribution of mutations, when calculating the percentage of synonymous substitutions [4, 5].

2.2 Mutation rate

The most common type of mutation is substitutions that give rise to single nucleotide polymorphisms (SNPs; in this thesis the term SNP refers to substitutions only and does not include single nucleotide deletions or insertions). The spontaneous rate of substitutions is estimated to be 7.2×10^{-11} per site per generation for *P. aeruginosa* [6];

however, the rate is dependent on several factors such as physiological factors (*e.g.* mutagens or stress responses) and the efficiency of DNA repair systems [7-10].

As substitution mutations appear stochastically and typically randomly across the genome, the mutation rate should be calibrated to supply enough beneficial mutations, but at the same time not too many detrimental mutations. If the mutation rate is too low, the population may not generate the genetic diversity necessary for an efficient exploration of novel phenotypic solutions. However, a high mutation rate also increases the number of detrimental mutations that will be a burden to the population (also known as the *genetic load*).

Two key factors that should be taken into account, when inferring the optimal mutation rate, are the population size and the strength of selection for beneficial mutations relative to the selection against deleterious mutations. Together with the mutation rate the population size determines the mutation supply rate, which is defined as the product of the mutation rate and the population size. As beneficial mutations are rare, a low mutation rate and a small population size will increase the amount of time it takes for a beneficial mutation to arise. As time becomes a limiting factor, it might be advantageous to increase the mutation rate on the expense of a higher genetic load. Accordingly, the direct selection of rare beneficial mutations can hereby indirectly promote the selection, *i.e.* a second-order selection, for alleles increasing the mutation rate, also referred to as hypermutator alleles [3, 11, 12].

2.3 Genetic adaptation

Adaptation of an organism to changing environmental conditions can be facilitated by either phenotypic acclimation or genetic adaptation.

Phenotypic acclimation is the change of phenotype in response to environmental cues without any inheritable genetic change. Nonetheless, the capacity of an organism to produce different phenotypes is limited, and phenotypic acclimation will therefore be inadequate for long-term adaptation of an organism to novel and permanent changes in its environment [13].

Genetic adaptation is the process by which adaptation is accommodated by selection of genetic variants with improved fitness in the environment. Hence, the success of

this mechanism to facilitate adaptation is dependent on the organism's (or a population of organisms') ability to generate adaptive genetic diversity. Natural selection will then act to increase the frequency of genetic variants giving rise to phenotypes with improved fitness in the given environment. If the imposed selection pressure remains constant, the beneficial genetic variant will eventually become dominant in the population and outcompete all less fit genetic variants. This phenomenon illustrated in **Figure 2** is known as a *selective sweep* because all other genetic variants are swept away causing a reduction in genetic diversity [14]. The time for a selective sweep to happen depends on the strength of the selection for the beneficial genetic variant, but may also be prolonged from the interference from other beneficial genetic variants that arises at approximately the same time. This phenomenon is known as *clonal interference* and slows down the overall rate of genomic evolution.

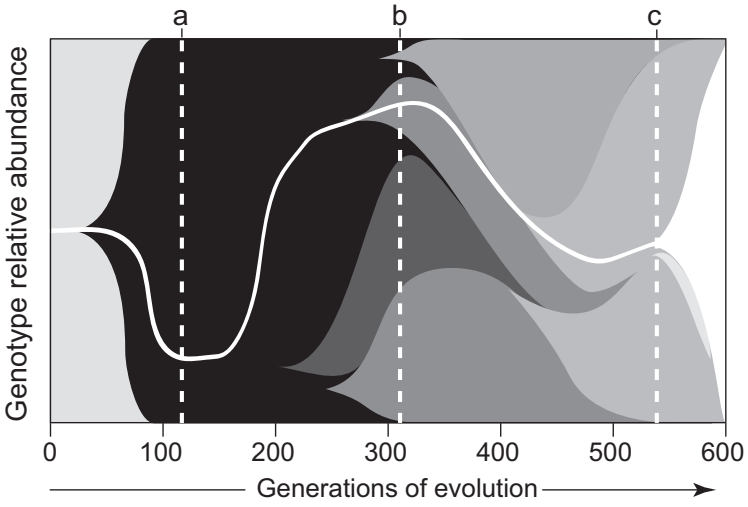


Figure 2: Genotype abundances in an evolving bacterial population. Lineages with new beneficial mutations are depicted as shaded wedges that originate in a previous genetic background and rise in frequency as they outcompete their ancestor and other lineages. The same shading indicates lineages that have equivalent fitness, and the path to the final dominant genotype containing five mutations is highlighted by the white curve. (a) A new beneficial mutation sweeps to fixation and the population has little diversity. (b) Four lineages with different mutations coexist at appreciable frequencies for a time before (c) the descendants of one lineage become a majority. Figure adapted from Barrick and Lenski (2009) [15].

2.4 Adaptation to heterogeneous environments

According to the above theory, only a single genotype will eventually occupy the sole niche offered by a constant and perfectly homogeneous environment [16]. This principle of *niche exclusion* is however not in effect when the environment shows either spatial or temporal heterogeneity.

Heterogeneity allows for the presence of multiple niches within the environment [17]. This means that, when a population evolves in a heterogeneous environment, it may genetically diversify to establish sub-lineages, each genetically adapting under the selection pressure of its niche. This process of genotypic diversification, for example shown for *Pseudomonas fluorescens* [18], is known as *adaptive radiation*.

The establishment of multiple niches and adaptive radiation is only possible because genetic adaptation towards one niche carries a cost of fitness in another niche. The cost may either be directly linked to natural selection of genetic variants that are beneficial in the given niche, but detrimental in other niches (a process known as *antagonistic pleiotropy*). Alternatively, the cost of adaptation to one niche may be due to nonadaptive accumulation of mutations that are neutral in the selective niche but detrimental in other niches.

Ecological specialization by adaptive radiation may, however, not always be effective in heterogeneous environments [17]. It may be the case in an environment with a temporal change, so the population needs to adapt to become a *generalist* with the highest possible average fitness (even if temporary suboptimal), rather than evolving into *specialist* populations with only temporary optimums of fitness.

2.5 Molecular phylogenetics

The previous sections introduced the fundamental evolutionary drivers that have made it possible for life to evolve in a step-wise manner to diversify and adapt to almost every possible environment on Earth. Nonetheless, as evolution is a slow process (compared to the lifespan of humans), we are impaired in our ability to follow the molecular evolution of complex traits as it happens. This section will serve to explain how phylogenetics can be applied to “rewind the tape of evolution” in order

decipher individual molecular steps and the genetic relationship of different organisms.

Molecular phylogenetics uses hereditary molecular differences (mainly in DNA sequences) to make inference about the evolutionary history of organisms. The most common approach is to compare homologous DNA sequences from different organisms. Assuming that the sequences of the alignment shares a common ancestor, it is possible to group organisms based on their similarity and to make hypotheses about ancestral states shared by different groups of taxa.

This cladistics approach to infer phylogeny assumes a vertical inheritance of genetic material, *i.e.* passage of genetic material from parent to offspring. This assumption is conflicted by HGT that confounds phylogenetic analysis, as HGT gives rise to genomes composed of genetic material with different phylogenies. In addition, phylogenetic analysis is also confounded by direct genetic reversion and convergent evolution in the same site as such instances may cause homoplastic sites (**Figure 3**).

Phylogenetic inference can be made on all types of genetic variation. Nonetheless, phylogenic inference based on SNPs in homologous DNA sequences is widely used for several reasons including that SNP mutations occur frequently, SNPs are typically random, and SNPs can be unambiguously defined in a sequence alignment. Furthermore, only a limited number of base substitutions exist for a given position in the genome (A/T↔G/C, A/T↔T/A, A/T↔C/G), and the rates of different substitution mutations can therefore be readily modeled. This is in contrast to insertion and deletion mutations that are unambiguous and define an infinite number of mutation types (*e.g.* with respect to number of nucleotides inserted/deleted).

In this thesis two different methods were used for phylogenetic inference: (a) maximum-parsimony, and (b) Bayesian analysis.

The optimality criterion for the maximum-parsimony method is to explain the observed data with the fewest possible number of evolutionary events (such as mutations) (**Figure 3**) [19]. This makes maximum-parsimony a simple approach not based on an explicit evolutionary model of substitution rates. Nonetheless, this means maximum-parsimony has no correction for multiple mutations at the same site, and it

makes it sensitive to differences in branch lengths or unequal rates of evolution in different lineages (leading to long branch attraction; **Figure 4**) [20].

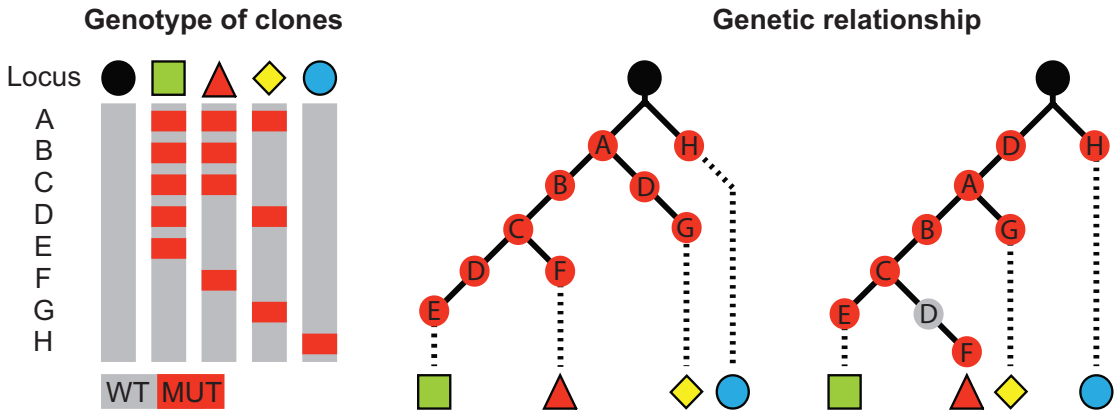


Figure 3: Maximum-parsimonious tree construction and homoplasy. Examples of two equally good maximum-parsimonious trees based on eight polymorphic loci (A-G). Both trees are optimized to explain the observed polymorphisms with the fewest possible number of mutational events. The occurrence of polymorphisms in the D locus is homoplastic, which means more than a single evolutionary event is needed to explain the phylogeny (either by convergent evolution or by direct reversion to the ancestral state). Homoplastic loci may lead to equally good maximum-parsimonious phylogenetic trees, and it such cases the phylogeny may be represented by a consensus of the different tree topologies.

As a probabilistic alternative to maximum-parsimony, Bayesian analysis is a parametric approach that evaluates the phylogeny with reference to an explicit evolutionary substitution model. The model allows for assigning parameters to describe the relative rate at which one nucleotide substitutes another during evolution, *e.g.* the relative rate of transitions to transversions. This means that for a given set of parameter values one can compute the probability of the observed sequence data, and the goal of Bayesian analysis is to obtain a full probability distribution over all possible parameter values. This again allows to make inference about phylogeny and individual parameters such a mutation rates or divergence times, when applying prior information on sampling times [21].

Overall, the intent of maximum-parsimony is to find the tree(s) that require the fewest steps, whereas Bayesian analysis also tries to minimize a function describing the amount of evolutionary change, but the priority is the likelihood of trees given an

evolutionary substitution model (and other prior information). The two methods may therefore produce different phylogenies. Nonetheless, in this thesis the two methods are not used for comparative reasons, but rather for separate purposes.

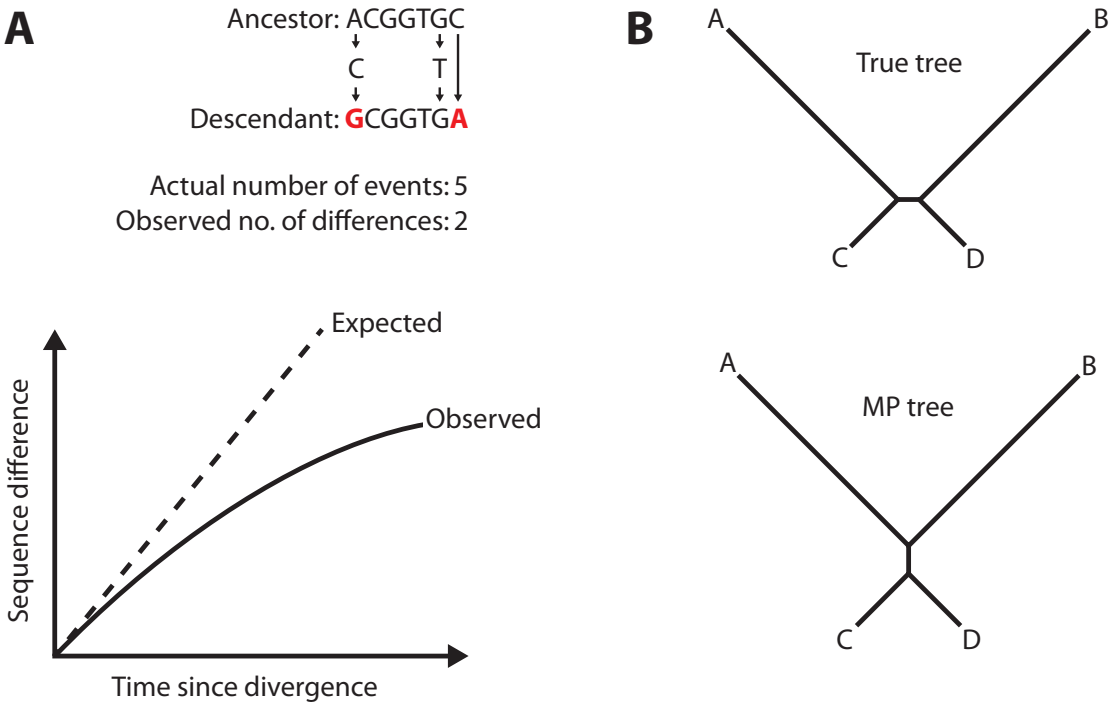


Figure 4: Superimposed substitutions and long branch attraction. (A) Comparison of homologues DNA sequences may not show all mutational event that have occurred since divergence, because the same site might mutate multiple times. The observed number of differences does therefore not scale linearly with the total evolutionary separation between sequences. For this reason, evolutionary substitution models have been designed to estimate the number of unseen mutational events to estimate corrected evolutionary distances. (B) Long branch attraction is a phenomenon that causes taxa to be clustered together due to the misinterpretation of shared homoplasies. The problem occurs when two taxa that have accumulated relatively many SNPs, by chance exhibits convergent evolution (*i.e.* they have accumulated the same SNPs independently). Maximum-parsimony (MP) may erroneously interpret the SNP as a single event (*i.e.* as a synapomorphy) in the common ancestor of the two taxa, if there are not other SNPs to outweigh the misinterpretation and support the correct ancestry. The problem may be minimized by using methods that correct for multiple substitutions at the same site, or by adding additional taxa to outweigh homoplastic SNPs.

Chapter 3

Pathoadaptation of *P. aeruginosa* to the cystic fibrosis airways

The most comprehensive way to acquire information about the genetic relationship between organisms is to sequence their entire genomes (*i.e.* phylogenomics). Historically the cost of DNA sequencing has been a limiting factor, so it has been prioritized to sequence members of genetically distant organisms to avoid spending valuable sequencing resources on re-sequencing of nearly identical genomes [22]. However, the advent of next generation sequencing (NGS) techniques has made it more affordable to sequence the genomes of more closely related organisms [23-25], and it is now feasible to follow the genomic evolution of bacteria by sequencing clones that have evolved for only a few generations since divergence. Such studies can easily be performed *in vitro* [4, 26-29], but even natural systems allows for sampling and re-sequencing of the same clonal lineage over shorter and longer time-periods [5, 6, 30, 31].

These approaches allow us to follow the microevolution of organisms and can improve our understanding of the selection pressures and mechanisms that govern genetic adaptation in specific settings. In the lab, this means that researchers can carry out evolution experiments to elucidate the genetic basis of adaptation of an organism to specific conditions, and for example use this information to genetically engineer organisms for industrial applications, or to build conclusions that can be extrapolated to help understand evolution in natural systems. However, it is not clear to what extent the laboratory observations relate to natural systems, why investigations of microevolution in complex and diverse natural environments are needed to fully understand evolutionary processes.

3.1 Clinical collections of bacteria as sources of insight into microevolution

The study of microbial evolution in natural systems is limited due to the difficulties of systematic and consistent sampling. For example, the constant change and progression of natural environments makes it difficult to obtain multiple samples from the same

niche, and the populations of interest may be small. Nonetheless, the systematic sampling of pathogenic bacteria from clinical settings has shown to provide researchers with an opportunity to study microevolution within the well-defined boundaries and tissues of the human host [6, 30, 32]. Furthermore, insight into molecular evolution of pathogenic bacteria is not only interesting from the evolutionary biologist's point of view, but the same knowledge is relevant from a clinical perspective in order to understand for example epidemiology and evolution of antibiotic resistance [33-38].

3.2 The cystic fibrosis model

The opportunistic pathogen *P. aeruginosa* is a common environmental inhabitant, which is also capable of causing both acute and chronic infections in a range of hosts from amoeba and plants to humans. For example, *P. aeruginosa* causes chronic airway infections in most patients with cystic fibrosis (CF), and it is directly associated with the morbidity and mortality connected with this disease. The microbiological status of CF patients is therefore followed systematically, and *P. aeruginosa* isolates are sampled at a routine basis when the patients visit the clinic. Thereby chronic CF infections provide an opportunity for long-term monitoring of the battle between the infecting bacteria and the host immune defense and clinical intervention therapy, and thus offer a direct method for observing evolutionary mechanisms *in vivo* [6, 39, 40].

3.2.1 Cystic fibrosis

Cystic fibrosis is the most common life-limiting inherited genetic disease among the Caucasians. Its frequency within this population group is 1 in 2,500 live births [41], and approximately 70,000 people per year are diagnosed with CF worldwide. The disorder is autosomal recessive, and results from the combination of two parental mutant alleles of the CFTR gene encoding the CF transmembrane conductance regulator (CFTR) [42]. More than 1,500 different mutations have been associated with CF; however, 70% of all CF patients carry the ΔF508 mutation. Defects in CFTR result in abnormal ion transport over epithelium surfaces leading to production of thick dehydrated mucus in the airways. The viscosity of the mucus impairs the

function of cilia to keep the airways free of inhaled material, and as a result CF patients are more susceptible to infections from inhaled microbes [43].

3.2.2 Cystic fibrosis airways pathogens

A range of organisms is found to cause infection in the airways of CF patients, and the species most frequently associated with infection of CF airway infections are *P. aeruginosa*, *Staphylococcus aureus*, *Haemophilus influenza*, and *Burkholderia cepacia* complex. In addition, culture-independent investigations have shown the diversity of bacteria in CF airways to comprise hundreds of other species [44-46], but the relevance of these findings is disputed [47].

The infection from different species is observed to be age-dependent. *S. aureus* and *H. influenzae* are the dominant species to cause infection in childhood, whereas *P. aeruginosa* is most frequently found pathogen in adults (**Figure 5**). Accordingly, about 80% of adults with CF have chronic *P. aeruginosa* infection, and thereby *P. aeruginosa* is major contributor to morbidity and mortality in CF patients.

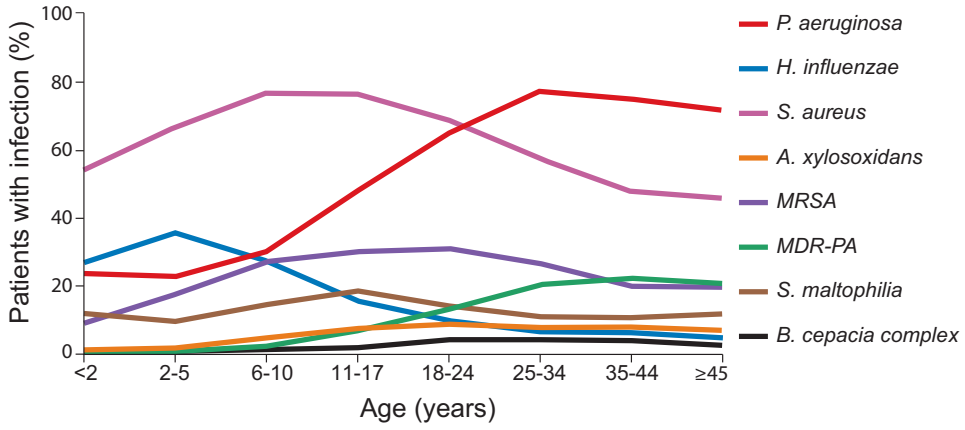


Figure 5: Prevalence of common pathogens in the airways of CF patients as a function of age. Organism abbreviations: *A. xylosoxidans*, *Achromobacter xylosoxidans*; *B. cepacia*, *Burkholderia cepacia*; *H. influenzae*, *Haemophilus influenzae*; MDR-PA, multidrug-resistant *P. aeruginosa*; MRSA, methicillinresistant *S. aureus*; *S. aureus*, *Staphylococcus aureus*; *S. maltophilia*, *Stenotrophomonas maltophilia*. *P. aeruginosa* includes patients with MDR-PA. *S. aureus* includes patients with MRSA. Figure adapted from the Cystic Fibsosis Foundation, Patient Registry Annual Data Report 2011 (www.cff.org).

It is not clear why *P. aeruginosa*, despite only being classified as an opportunistic pathogen, in the long run is a more successful colonizer of CF airways than pathogens like *S. aureus* and *H. influenza*. One possibility is that the genetic repertoire of *P. aeruginosa* inherits a better potential for genetic adaptation to the CF airways. The following paragraphs serves to introduce the environment of the CF airways and *P. aeruginosa* itself.

3.2.3 The environment of the CF airways

The CF airways constitute a complex environment with both spatial and temporal heterogeneity (**Figure 6**). In the upper airway compartments *P. aeruginosa* is found to colonize the paranasal sinuses causing sinusitis [48], whereas *P. aeruginosa* in the lower airways is mainly found in the viscous mucus layer of the bronchi [49]. In the respiratory zone *P. aeruginosa* is more rare, except in the cases of severe lung damage [50].

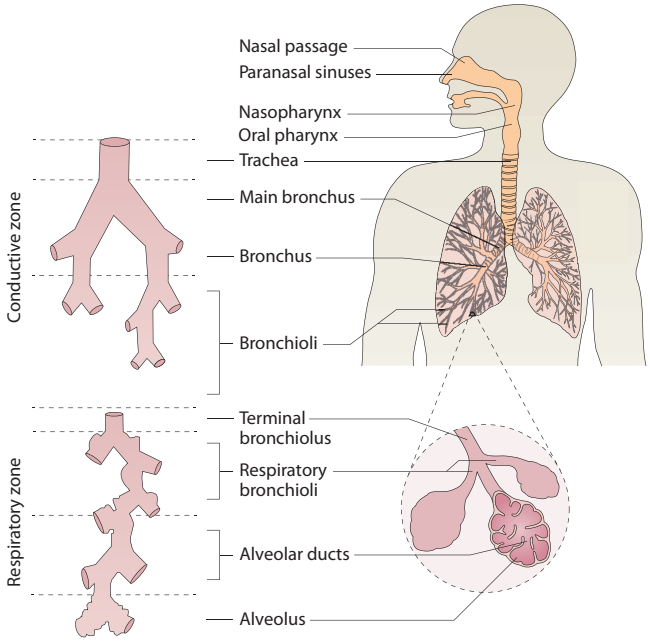


Figure 6: The human airways. The upper airways include the nasal passage, the paranasal sinuses, pharynx, and larynx (can also be considered as part of the lower airways). The lower airways consist of the conductive zone and the respiratory zone. The conductive zone refers mainly to the bronchi, and the respiratory zone comprises the respiratory bronchioles and the alveoli. Figure adapted from Folkesson *et al.* (2012) [49].

Within the CF airways, *P. aeruginosa* is faced to adapt to several stresses and conditions that stand out from the environment outside the host. Among others, this includes the immune system, antibiotics, oxygen levels, and nutrient availability.

Immune system

The immune system is considered as one of the most significant challenges for microbes invading the CF airways. The immune response includes (aside from the impaired mucociliary escalator) components such as neutrophils, macrophages, defensins, and antibodies. The type and level of response may be dependent on the compartment of the airways. For example the response from neutrophils is more predominant in the lower airways than in the sinuses [51]. Neutrophils and macrophages suppress infection by phagocytosis of bacterial invaders; however, in addition to this invaders need to survive oxidative stress from phagocytosis-mediated liberation of reactive oxygen species (ROS) [52]. Nonetheless, ROS also cause collateral damage to the host tissues which leads to the loss of lung function [53]. Furthermore, part of the immune response relies on the recognition of microbial epitopes (*e.g.* in lipopolysaccharide or flagellar epitopes), and the invading bacteria may therefore adapt to the airways by reducing their immunogenicity [54].

Antibiotics

The antibiotic treatment regime of CF patients varies between countries and clinics [55]. In the Copenhagen CF clinic a broad range of classes of antibiotics is used (*e.g.* aminoglycosides, beta-lactams, antimicrobial peptides, macrolides, and fluoroquinoles), and treatment depends on the bacteriology and the condition of the individual patient [56-58]. A combination of both oral and intravenous administration can be used to obtain effective concentrations in different compartments of the lung, but the distribution and dynamics of the different drugs will inevitable vary in different compartments of the airways [59, 60].

Oxygen levels

One might think that the human airways constitute a highly aerated environment; however, both aerobic and anaerobic zones exist. The respiratory zone, where the gas-exchange occurs, is generally aerobic while oxygen levels in the mucus-enriched

zones (conductive zone and paranasal sinuses) can range from aerobic over microaerobic to anerobic due poor oxygen diffusion into the mucus [61, 62]. Nonetheless, the mucus contains sufficient amount of alternative electron acceptors, such as nitrate, to facilitate anaerobic respiration by facultative anaerobes like *P. aeruginosa* [63].

Nutrients, iron, and salts

The CF airways are considered as a nutrient-rich environment. Mucus is reported to be particular rich on amino acids, but also other energy sources such as glucose, lactate, and different types of lipids are abundant [64, 65]. Despite the apparent nutrient richness, considerable metabolic adaptation from bacterial invaders may still be essential for successful persistence in the airways [66, 67].

As scavenging of iron is a hallmark of the battle between pathogens and their hosts, the iron availability in the CF airways is of interest [68]. The hosts iron-withholding processes may limit the amount of free iron, what makes *P. aeruginosa* dependent on siderophores and heme uptake systems to scavenge iron from the host [69]. Nonetheless, uptake of free available ferrous iron by *P. aeruginosa* is also important as infections progresses and host tissues are deteriorated from inflammation [70, 71].

While iron may be a scarce source, salts (Na^+, K^+, and C^-) are on the other hand present in high concentrations due to the abnormal ion transport caused by the malfunction of CFTR [72, 73]. This means invading bacteria may need to cope with high osmotic pressures in order to survive [74].

3.3 Pseudomonas aeruginosa

The focus of this thesis is to describe the molecular evolution and genetic adaptation of *P. aeruginosa* within the CF airways described above. The following section will therefore serve to describe *P. aeruginosa* with respect to its the genetic repertoire and the most important CF-related phenotypes of *P. aeruginosa*.

P. aeruginosa is a Gram-negative bacterium found in a wide variety of ecological niches in soil, marshes, and marine habitats. It rarely causes infection in healthy humans, but is able to infect damaged tissues or those with reduced immunity, and it is for example found in found in skin burns, urinary tract infections, and in the lungs

of CF patients [75]. The ability of *P. aeruginosa* to thrive in a diverse range of environments is proposed to be due to a relative large genome size encoding a high proportion of regulatory proteins [76].

The 'Pseudomonas Genome Database' [77] currently contains 13 public available genome sequences of *P. aeruginosa* ranging in size from 6.2 to 6.9 mega base pair (Mbp) and with GC contents of around 66%. The genome is most extensively annotated with reference to *P. aeruginosa* strain PAO1 (genome size 6.3 Mbp with 5,570 predicted open reading frames), which shares 94.7% of its genetic content with the strain in focus in this thesis, *P. aeruginosa* strain DK2 [39].

3.3.1 Pseudomonas aeruginosa infections in cystic fibrosis

The ability of *P. aeruginosa* to cause infection in CF patients is not pertained to a single or a few strains (clone types) of *P. aeruginosa*. In contrast, many different and unique strains are observed to infect CF patients [78]. This suggests that patients pick up *P. aeruginosa* from diverse environmental sources, and accordingly multiple *P. aeruginosa* strain types are typically found to cause intermittent infections in the early lives of the individual patient. However, the frequency, duration and severity of infections will, despite clinical intervention therapy, increase and as the patient regresses into a chronic infection stage, a single strain type is typically observed to be dominant (**Figure 7**) [49].

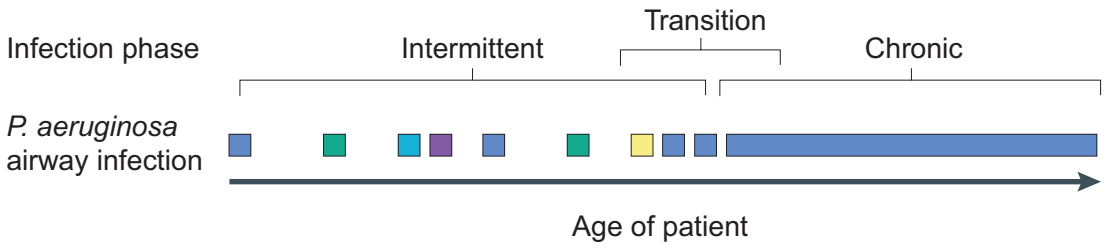


Figure 7: Clonal diversity and progression *P. aeruginosa* infections in cystic fibrosis patients. Colored squares represent the detection of *P. aeruginosa* in the airways of a hypothetical CF patient. The different colors of the squares symbolize that different *P. aeruginosa* clone types may be found to cause infection. Polyclonal infection patterns are most often found in the early stages of the infection, where after a single clone type is observed to dominate. Re-appearance of the same clone type after

periods with no detectable *P. aeruginosa* might either be due to re-colonization from the same environmental source, or from hidden reservoirs within the patient [79, 80].

3.3.2 Adaptive phenotypes of Pseudomonas aeruginosa

Genetic adaptation is found to play a major role in the successful establishment of chronic *P. aeruginosa* infections of CF patients [5, 6, 39, 40, 81, 82]. Accordingly, it is observed that natural selection acts on *P. aeruginosa* in the CF airways to accommodate the fixation of mutations causing beneficial phenotypic changes. These mutations are here referred to as *pathoadative mutations*.

The strains isolated from early and intermittent infections have phenotypes like *P. aeruginosa* found in the environment. They are characterized by being non-mucoid, fast growing, and relatively susceptible to antibiotics [83]. However, the phenotypes of *P. aeruginosa* isolates are observed to change together with the transition into the chronic infection stage, and isolates from chronic infections are observed to form mucoid colonies, reduction in expression of virulence factors, loss of flagella dependent motility, amino acid auxotrophy, hypermutability, antibiotic resistance, and various lipopolysaccharide (LPS) modifications [84-90]. The genetic basis and characteristics of phenotypes related to mucoidity, antibiotic resistance, loss of virulence factors, and hypermutation are explained below.

Mucoidity

The mucoid phenotype is characterized by overproduction of the exopolysaccharide alginate resulting in a slimy colony morphology when grown on agar plats. It is proposed that the alginate functions as a barrier to protect *P. aeruginosa* from both the immune system response and exposition to antibiotics [91-93]. This might explain why the mucoid phenotype is considered as a hallmark of chronic infection, and it is associated with poor prognosis of CF patients [94]. Constitutive mucoidity is genetically most often caused by loss-of-function mutations in *mucA* which encodes an anti-sigma factor, that sequesters (represses) the sigma factor AlgT (sigma22) [95]. AlgT controls the expression of a range of stress response genes associated with the cell envelope, including the genes required for alginate production.

Antibiotic resistance

Development of increased resistance towards antibiotics is another trait that is commonly observed for *P. aeruginosa* from chronic CF infections. *P. aeruginosa* is generally characterized as being relative resistant to many classes of antibiotics due to a low outer membrane permeability, the production of an AmpC beta-lactamase, and the expression of multiple drug efflux pumps [96]. The selection pressure enforced by antibiotic treatment is therefore hypothesized to promote CF airway colonization opportunities for *P. aeruginosa* as *S. aureus* was the dominating CF pathogen until effective antibiotic therapies were implemented [49]. Nonetheless, despite a high intrinsic tolerance or resistance to many antibiotics, *P. aeruginosa* is still observed to develop and enhance its antibiotic resistance by accumulation of mutations.

Commonly observed resistance mutations are: (a) mutations that increase the activity of drug efflux pumps (*e.g.* drug efflux system MexAB-OprM and MexXY-OprM [97, 98]); (b) loss-of-function mutations in membrane porine OprD leading decreased influx of carbapenems [99]; (c) mutations that increase the resistance towards cationic antimicrobial peptides by affecting the composition of the LPS [100-102]; (d) mutations that alter the direct targets of antibiotics (*e.g. gyrA* encoding a DNA gyrase targeted by fluoroquinolones [103, 104]); and (e) mutations that increase the activity of beta-lactamases [105].

Loss of virulence factors

P. aeruginosa is able to produce a number of virulence factors which are reported to contribute to its pathogenesis. This includes cell-associated virulence determinants such as flagella, type IV pili, and LPS, and extracellular virulence factors such as siderophores, exotoxins, pyocins, and proteases. Note however, that the role and importance of virulence factors is dependent on the infection context, and the virulence of the above-mentioned factors is mostly defined with respect to acute infection models. Accordingly, it is less counter-intuitive that *P. aeruginosa* from chronic infections is frequently observed to have lost the ability to express virulence factors [84, 89, 106, 107]. This loss of virulence factors is hypothesized to be due either to a selection against the production of immunogenic determinants, or a selection for minimization the energy cost of production of virulence factors [54].

Furthermore, the loss of virulence factors have been found to be caused by genes encoding regulators and sigma factors such as Vfr, LasR, RpoN, AlgT, and PvdS [5, 66, 74, 84].

Hypermutation

In year 2000 Oliver and colleagues reported a high proportion of *P. aeruginosa* isolates from chronic CF infections to exhibit increased rates of mutation [88]. The hypermutable phenotypes were most commonly caused by loss-of-function mutations in the *mutS* and *mutL* genes encoding components of the DNA mismatch repair system, and it was suggested that elevated mutation rates may confer a selective advantage in the CF airways. Faster acquisition of for example antibiotic resistance mutations may be one explanation for the high prevalence of hypermutators in CF infections; however, the exact adaptive advantage from an elevated mutation rate remains unclear [108-111]. For example, the hypermutators may exhibit a particular mutation bias (mutational signature) that effect not only the rate, but also the genetic mechanism of adaptation, as it has been shown for the biased mutagenesis of *mucA* in *mutS* hypermutators [112]. Another possible explanation is that hypermutators occupy a niche, in which there is a selection for fast phenotypic changes, and as such both hyper- and normomutators would be able to co-exist in the airways.

Chapter 4

Present investigations

Chapter 2 has taught us how phylogenomics can be used to decipher the molecular evolutionary steps that have occurred since the divergence of compared organisms, such as isolates of bacteria. However, the retrospective nature of phylogenomics makes the approach dependent on the quality of the collection of bacteria that is of interest. Ideally, samples representing all genotypic states should be contained in the collection. In practice this is not feasible, and especially lack of information about of ancestral states is a problem because samples representing these states are in most cases not accessible at the time when the study is conducted.

Nonetheless, freeze-storage of bacteria provides an opportunity to gain insight to the intermediates of evolution. Such freeze-storage of ancestral states have for example founded the basis of the longest running laboratory evolution experiment in which Lenski and colleagues have followed the evolution within each of 12 test tube cultures of *Escherichia coli*, that have been propagated since year 1988 (corresponding to >50,000 bacterial generations) [113]. In a similar way, the systematic sampling and freeze-storage of clonal lineages of pathogenic bacteria from clinical settings provide collections that can be regarded as 'evolution experiments', but with the notable distinction that these are run in natural settings.

Chapter 3 then taught us how chronic *P. aeruginosa* infections in CF patients represent an infectious disease scenario, in which the molecular evolution of clonal lineages of *P. aeruginosa* within the airways of CF patients can be followed. It is therefore in this context that this thesis aims to improve our knowledge about molecular evolution and pathoadaptation, and our investigation is based on a unique collection of *P. aeruginosa* isolates sampled from Danish CF patients.

4.1 The Copenhagen collection of *Pseudomonas aeruginosa*

The collection of clinical isolates of *P. aeruginosa* sampled from CF patients dates back to 1972, where it was started by Professor Niels Høiby at the Copenhagen CF Center at Rigshospitalet in Denmark. Since then, *P. aeruginosa* have been isolated and stored on a regular basis from patients attending the Copenhagen CF Center, and today thousands of isolates are stored.

Molecular epidemiological studies of selected isolates in the collection revealed the presence of two dominant strains that have spread among the CF patients by interpatient transmission [80]. The two strains are referred to as the DK1 (formerly "r") and the DK2 (formerly "b") clone type, respectively. The earliest sampled isolates of the DK1 and DK2 clone types were sampled in 1973 and 1972, respectively, and since then they been able to disseminate through a cohort of around 40 patients each (**Figure 8**). This means that it possible to follow the genomic evolution of these two clone types over a time-period covering more than 40 years, corresponding to around 200,000 bacterial generations [6]. This is in the range of the number of generations separating present-day humans and chimpanzees from their most recent common ancestor [114].

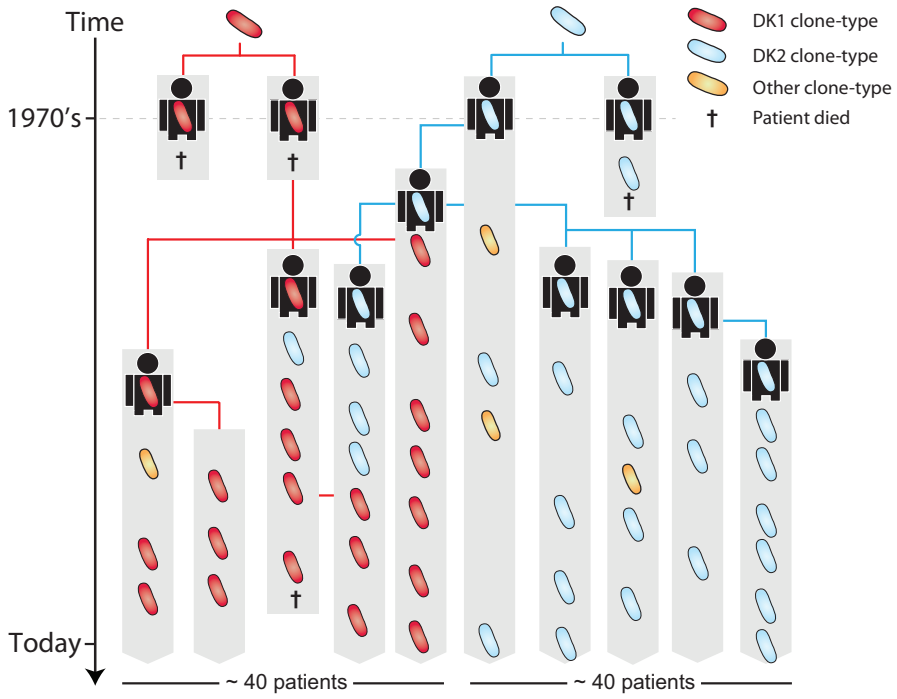


Figure 8: Schematic representation of the epidemiology of the *P. aeruginosa* DK1 and DK2 clone types causing infections in Danish CF patients. Illustration of the initial entry and spread of the DK1 and DK2 clone types at the Copenhagen CF Center. Figure courtesy of Søren Damkiær.

Members of the DK1 and DK2 clone types were initially identified from genotyping using pulsed field gel electrophoresis (PFGE) and SNP typing based on 13 polymorphic sites in the *P. aeruginosa* core genome (Clondiag Chip Technologies, Germany) [115]. Neither of these typing methods is able to provide details about microevolutionary events, so we started a program to genome sequence selected isolates in the Copenhagen CF collection of *P. aeruginosa*. To this date more than 600 isolates belonging to more than 50 different clone types have been genome sequenced with the DK1 and DK2 clone types being the most extensively investigated.

4.2 Aim of thesis

Given this unique collection of *P. aeruginosa* and the ability to do comparative genomics, the aim of this thesis has been to use the CF model to study evolution and pathoadaptation of bacteria in the course of persistent infections in humans. This has been done with a focus on the DK2 clone type, and the articles presented in Chapter 6 aim to investigate evolution within the CF airways with respect to:

- Rates and mechanisms of molecular evolution.
- Population dynamics.
- Identification of pathoadaptive genes and selective forces.

4.3 Overview of results

Article 1: Evolutionary dynamics of bacteria in a human host environment

Our first efforts to use genome sequencing to investigate the molecular evolution of *P. aeruginosa* was presented in **Article 1** by Yang *et al.* (2011) [6]. In this article we compared the genomes of 12 DK2 isolates sampled from six different patients over 35 years. Phylogenomic analysis revealed the evolution of the DK2 lineage to be characterized by a rather constant accumulation of SNPs exhibiting an overall signature of genetic drift (dN/dS=0.8), and only 180 SNPs separated the two most distant isolates (CF114-1973 and CF333-2007a).

Only the branch separating the 1973-isolates from the post-1979 isolates showed significant signs of positive selection (dN/dS=1.6), suggesting that the DK2 lineage during this period experienced a more rapid acquisition of adaptive mutations. Interestingly, phenotypic characterization of the 12 isolates showed a rapid phenotypic change during the same period. We therefore hypothesized that the DK2 lineage by 1979 had reached an adaptive peak in the fitness landscape mediated by the fixation of important pathoadaptive mutations. The presence of such pathoadaptive mutations fixated before 1979 was further investigated, and Damkiær *et al.* (2013) found four specific mutations in genes *mucA*, *algT*, *rpoN*, and *lasR* to explain 40% of the total transcriptome changes, to cause conditional mucoidity, and to increase the tolerance to towards relevant antibiotics [74].

The topology of the phylogenetic tree describing the genetic relationship of the 12 isolates showed that isolates from the same patient tended to group together. This suggested that the DK2 lineage subsequent to transmission evolves as rather separate sub-lineages in each individual patient. As the selective conditions in the different hosts may be similar, the relationship between genotype and phenotype can be elucidated from parallelism in evolution, *i.e.* independent evolution of the same phenotype from mutation of the same locus. Accordingly, we were able to infer the genetic details of two phenotypes (increased resistance towards ciprofloxacin and lost capacity to catabolize 4-hydroxyphenylacetic acid) by comparing genotypic and phenotypic changes in independently evolving sub-lineages.

Article 2: Mutations in 23S rRNA confer resistance against azithromycin in Pseudomonas aeruginosa

In **Article 2** by Marvig *et al.* (2012), we characterized another example of convergent evolution in which isolates CF333-2007a and CF66-2008 sampled from two different patients were found to have independently accumulated the same mutation within 23S rRNA [116]. This led us to reveal how mutations in domain V of 23S rRNA confer resistance towards the macrolide antibiotic azithromycin. Azithromycin has been used in the Copenhagen CF Center to treat *P. aeruginosa* CF infections since 2001 [56]; however, our study is the first to demonstrate *in vivo* development of resistance in *P. aeruginosa*. Furthermore, the exact mechanism by which azithromycin exerts its

effect has not been well understood, and our results suggest that the mode of action of azithromycin *in vivo* involves binding to the 50S ribosomal subunit, and that the action is blocked by mutations in 23S rRNA.

Article 3: Deletion and acquisition of genomic content during early stage adaptation of Pseudomonas aeruginosa to a human host environment

Our initial genome analysis of the DK2 lineage only included a systematic detection of SNP mutations. Nonetheless, the dynamics of the overall genomic content of the DK2 lineage is of interest to determine the role of loss and acquisition of genetic material for long-term adaptation to the CF airways. For example the lineage may acquire plasmid-borne antibiotic resistance genes, or on the other hand undergo reductive evolution through loss of genetic material that are not essential within the host environment. In **Article 3** by Rau *et al.* (2012), we therefore examined the dynamics and mechanisms of changes in genetic content during evolution of the DK2 lineage [39]. At the same time we expanded the genomic dataset from the 12 initial isolates, so it now included genome sequences of 45 *P. aeruginosa* isolates collected from 16 individuals over 35 years, and we completed the 6,402,658 bp genome of CF333-2007a to strengthen the genetic basis of the analysis. In total 27 deletion events that occurred through both illegitimate and homologous recombination were detected. On average the deletions contained 45 kilo basepair (kbp) of DNA, and the rate of genome reduction was nearly 4 kbp per year. On the other hand, we observed no strong evidence of uptake of novel DNA, demonstrating that host adaptation was characterized by a reduction of the genomic repertoire rather than acquisition of novel functions. Also, the impact on evolution from transposition of IS elements was limited, and only five multiplication events of IS elements were observed. This result contrast laboratory evolution experiments in which IS elements were shown to contribute significantly to mutation generation [4, 117].

Article 4: Genome Analysis of a Transmissible Lineage of Pseudomonas aeruginosa Reveals Pathoadaptive Mutations and Distinct Evolutionary Paths of Hypermutators

In **Article 4** by Marvig *et al.* (2013), we again expanded the genomic dataset to include genome sequences of a total of 55 DK2 isolates [40]. The expansion of the

dataset to include isolates from 21 different patients enabled us to perform an open search for pathoadaptive genes by tracking patterns of parallel evolution in separate hosts. Accordingly, we identified 65 pathoadaptive genes of which most were related to antibiotic resistance, the cell envelope, or regulatory functions, suggesting that selection from the immune system and drug therapy drives the molecular evolution. Furthermore, the sequencing of multiple isolates from the same patient gave us an opportunity to study the within-patient population dynamics, and we found evidence for all of the following scenarios: (a) stably existence of a single clonal sub-lineage; (b) stably co-existence of two independent sub-lineages, possible due to niche-differentiation; and (c) sequential dominance of competing sub-lineages within the same patient. In the latter case the prevalence of pathoadaptive mutations interestingly correlated with the evolutionary success of the competing sub-lineages.

Finally, we observed the presence of hypermutable sub-lineages within 48% of the studied patients. While hypermutation may accelerate evolution, our investigations also led us to show that hypermutators display differential mutagenesis facilitated by mutation of homopolymers. This enabled the hypermutators to follow alternative evolutionary pathways that may help to explain their emergence in CF infections.

Article 5: Within-Host Evolution of Pseudomonas aeruginosa Reveals Adaptation Towards Iron Acquisition from Hemoglobin

The identification of pathoadaptative genes in the previous article was based on only intragenic mutations. Nonetheless, natural selection also acts on mutations located outside genes, for example due to the role of intergenic sequences in regulation and transcription of neighboring genes. We therefore re-analyzed the data with respect to intergenic regions, and in **Article 5** we found the intergenic region upstream of the *phuR* gene to be the most densely mutated intergenic region with a total of 13 mutations [118]. *phuR* encodes an outer membrane receptor of the *Pseudomonas* heme uptake (*phu*) system, and this finding led us to elucidate how the mutations lead to an increased activity of the *phuR* promoter. Furthermore, increased expression of *phuR* conferred a growth advantage in the presence of hemoglobin, thus suggesting that *P. aeruginosa* genetically adapt towards iron acquisition from hemoglobin in the CF airways. To test the generality of this adaptive trait, we inspected the genomes of

two additional *P. aeruginosa* lineages isolated from CF airways and found similar adaptive evolution in both of these lineages. Finally, in all three *P. aeruginosa* lineages *phuR* promotor mutations coincided with the loss of pyoverdine production, suggesting that within-host adaptation towards heme utilization is triggered by the loss of pyoverdine production.

Chapter 5

Overall conclusions, future directions, and perspectives

By genome sequencing of 55 isolates of the DK2 lineage of *P. aeruginosa* sampled from CF patients, we have shown how a clinical collection of bacteria constitutes a valuable basis for studying the molecular evolution of bacteria in natural environment. We analyzed the genomes to identify SNPs, insertions, deletions, and transposition of IS elements, but other mutation types (*e.g.* rearrangements and gene duplications) remain to be investigated. Our genome comparisons revealed within-host evolution of *P. aeruginosa* to be characterized by *de novo* mutation rather than horizontal transfer of genetic material, and we were able to give unprecedented estimates of *in vivo* rates of mutation. In this respect, we found certain hypermutable lineages to evolve with different rates and spectra of mutations, and this may have had impact on the evolutionary trajectories that were available for these hypermutators. Furthermore, one of hypermutator lineages had accumulated >2000 SNPs since its divergence from the other DK2 isolates, highlighting the difficulty of imposing a simple threshold for the number of SNPs between isolates to decide whether they are part of a recent transmission chain.

We observed the occurrence of multiple different DK2 sub-lineages in individual patients, and this suggested that population dynamics were driven by both intra-clonal competition (clonal interference) and niche specialization (adaptive radiation). We therefore propose that multiple and randomly chosen isolates from each patient should be genome sequenced to give a more detailed view of the population dynamics, as our current study is limited in sense that only a single (or a few) isolates from the same sputum sample were sequenced. Also, our genome analysis may be complemented with culture-independent sequencing of the *P. aeruginosa* population (metagenomics), or with sequencing of clones sampled from different sites in the airways to add spatial resolution our the study.

By tracking patterns of parallel evolution, we identified genes that were mutated to optimize the fitness of *P. aeruginosa* within the CF airways. The identification of such pathoadaptive genes gave insight into how *P. aeruginosa* evolves under the selective pressures imposed by drug therapy and the host immune system. The role in pathogenesis is not known for several of the identified genes, and further investigation of these genes may identify them as future therapeutic targets against infection. For example, we found evidence that *P. aeruginosa* during later stages of infection looses the capability to scavenge iron by pyoverdine siderophores and instead depends on iron acquisition via the *phu* heme uptake system. Targeting heme utilization might therefore be a promising strategy for the treatment of CF infections.

Continued characterization of pathoadaptive mutations will help linking genotype to phenotype. Such knowledge will facilitate better epidemiological predictions and provide the clinicians with valuable information on how to treat and segregate patients. For example, information about resistance mutations can be used to predict antibiotic susceptibility profiles even before phenotypic susceptibility results from standard lab tests are available [33].

We found evidence that adaptation towards heme utilization might be a general adaptive trait across different CF associated *P. aeruginosa* lineages. Nonetheless, it remains unclear how the findings from the DK2 genomes apply to other lineages of *P. aeruginosa* evolving in CF patients, and we anticipate that our dataset will facilitate inter-clonal future comparative studies. This might be by comparison to transmissible lineages from other countries, but we will also be able to do comparisons to the other "local" clone types found in the Copenhagen CF Center. Such local comparisons might for example be interesting with respect to shedding new light on interactions between different clone types that co-exists in the same patient. Our current study already indicates the importance of such interactions as our analysis revealed a DK2 isolate to have acquired DNA from an unrelated clone type of *P. aeruginosa*.

Furthermore, one drawback of the studied collection of DK2 isolates is that all the isolates have been sampled from their respective hosts after the onset of chronic infection. In order to better understand the genetic adaptation that occurs right after the movement from the environmental reservoir to the CF airways, we anticipate that

similar studies should be conducted on very young CF patients that experiences infection from naïve environmental lineages of *P. aeruginosa*. For example, the identification of pathoadaptive mutations in our study was based on detection of parallel evolutionary events. Thereby, we were limited in our ability to identify mutations that are of immediate importance for *P. aeruginosa* upon transition from the environmental reservoir to the CF airways, as such events may only have occurred once in the history of the DK2 lineage.

Finally, infectious disease scenarios are not as simple as one host and one pathogen. A logical next step will be to integrate genomic information from multiple different co-infecting pathogens to allow for deciphering interspecies interactions. Also, sequencing the genomes of the infected individuals may lead to a better understanding of the microbial evolution.

Overall, in this thesis we were able to: (a) track the molecular evolution of bacteria in a natural environment; (b) identify the genes involved in the genetic adaptation to this environment; and (c) gain information about the different evolutionary trajectories that were available for the bacteria.

Bibliography

1. Mojzsis, S.J., et al., *Evidence for life on Earth before 3,800 million years ago.* Nature, 1996. **384**(6604): p. 55-9.

2. Yang, Z. and J.P. Bielawski, *Statistical methods for detecting molecular adaptation.* Trends Ecol Evol, 2000. **15**(12): p. 496-503.

3. Wielgoss, S., et al., *Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load.* Proc Natl Acad Sci U S A, 2013. **110**(1): p. 222-7.

4. Barrick, J.E., et al., *Genome evolution and adaptation in a long-term experiment with Escherichia coli.* Nature, 2009. **461**(7268): p. 1243-7.

5. Smith, E.E., et al., *Genetic adaptation by Pseudomonas aeruginosa to the airways of cystic fibrosis patients.* Proc Natl Acad Sci U S A, 2006. **103**(22): p. 8487-92.

6. Yang, L., et al., *Evolutionary dynamics of bacteria in a human host environment.* Proc Natl Acad Sci U S A, 2011. **108**(18): p. 7481-6.

7. Sniegowski, P.D., P.J. Gerrish, and R.E. Lenski, *Evolution of high mutation rates in experimental populations of E. coli.* Nature, 1997. **387**(6634): p. 703-5.

8. Reuven, N.B., et al., *The mutagenesis protein UmuC is a DNA polymerase activated by UmuD', RecA, and SSB and is specialized for translesion replication.* J Biol Chem, 1999. **274**(45): p. 31763-6.

9. Wagner, J., et al., *The dinB gene encodes a novel E. coli DNA polymerase, DNA pol IV, involved in mutagenesis.* Mol Cell, 1999. **4**(2): p. 281-6.

10. Tang, M., et al., *UmuD'(2)C is an error-prone DNA polymerase, Escherichia coli pol V.* Proc Natl Acad Sci U S A, 1999. **96**(16): p. 8919-24.

11. Mao, E.F., et al., *Proliferation of mutators in A cell population.* J Bacteriol, 1997. **179**(2): p. 417-22.

12. Tenaillon, O., et al., *Mutators and sex in bacteria: conflict between adaptive strategies.* Proc Natl Acad Sci U S A, 2000. **97**(19): p. 10465-70.

13. Via, S., et al., *Adaptive phenotypic plasticity: consensus and controversy.* Trends Ecol Evol, 1995. **10**(5): p. 212-7.

14. Burke, M.K., *How does adaptation sweep through the genome? Insights from long-term selection experiments.* Proc Biol Sci, 2012. **279**(1749): p. 5029-38.

15. Barrick, J.E. and R.E. Lenski, *Genome-wide mutational diversity in an evolving population of Escherichia coli.* Cold Spring Harb Symp Quant Biol, 2009. **74**: p. 119-29.

16. Hardin, G., *The competitive exclusion principle.* Science, 1960. **131**(3409): p. 1292-7.

17. Kassen, R. and P.B. Rainey, *The ecology and genetics of microbial diversity.* Annu Rev Microbiol, 2004. **58**: p. 207-31.

18. Rainey, P.B. and M. Travisano, *Adaptive radiation in a heterogeneous environment.* Nature, 1998. **394**(6688): p. 69-72.

19. Fitch, W.M., *Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology.* Systematic Zoology, 1971. **20**(4): p. 406-416.

20. Felsenstein, J., *Cases in which parsimony and compatibility methods will be positively misleading.* Systematic Zoology, 1978. **27**(4): p. 401-410.

21. Drummond, A.J. and A. Rambaut, *BEAST: Bayesian evolutionary analysis by sampling trees.* BMC Evol Biol, 2007. **7**: p. 214.

22. Parkhill, J., *What has high-throughput sequencing ever done for us?* Nat Rev Microbiol, 2013. **11**(10): p. 664-5.

23. Shendure, J., et al., *Accurate multiplex polony sequencing of an evolved bacterial genome.* Science, 2005. **309**(5741): p. 1728-32.

24. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.

25. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.

26. Palmer, K.L., et al., *Genetic basis for daptomycin resistance in enterococci.* Antimicrob Agents Chemother, 2011. **55**(7): p. 3345-56.

27. Linkevicius, M., L. Sandegren, and D.I. Andersson, *Mechanisms and fitness costs of tigecycline resistance in Escherichia coli.* J Antimicrob Chemother, 2013.

28. Wong, A., N. Rodrigue, and R. Kassen, *Genomics of adaptation during experimental evolution of the opportunistic pathogen Pseudomonas aeruginosa.* PLoS Genet, 2012. **8**(9): p. e1002928.

29. Ensminger, A.W., et al., *Experimental evolution of Legionella pneumophila in mouse macrophages leads to strains with altered determinants of environmental survival.* PLoS Pathog, 2012. **8**(5): p. e1002731.

30. Lieberman, T.D., et al., *Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes.* Nat Genet, 2011. **43**(12): p. 1275-80.

31. Denef, V.J. and J.F. Banfield, *In situ evolutionary rate measurements show ecological success of recently emerged bacterial hybrids.* Science, 2012. **336**(6080): p. 462-6.

32. Croucher, N.J., et al., *Rapid pneumococcal evolution in response to clinical interventions.* Science, 2011. **331**(6016): p. 430-4.

33. Koser, C.U., et al., *Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak.* N Engl J Med, 2012. **366**(24): p. 2267-75.

34. Harris, S.R., et al., *Whole-genome analysis of diverse Chlamydia trachomatis strains identifies phylogenetic relationships masked by current clinical typing.* Nat Genet, 2012. **44**(4): p. 413-9, S1.

35. Sun, G., et al., *Dynamic population changes in Mycobacterium tuberculosis during acquisition and fixation of drug resistance in patients.* J Infect Dis, 2012. **206**(11): p. 1724-33.

36. McAdam, P.R., et al., *Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant Staphylococcus aureus.* Proc Natl Acad Sci U S A, 2012. **109**(23): p. 9107-12.

37. Mutreja, A., et al., *Evidence for several waves of global transmission in the seventh cholera pandemic.* Nature, 2011. **477**(7365): p. 462-5.

38. Morelli, G., et al., *Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity.* Nat Genet, 2010. **42**(12): p. 1140-3.

39. Rau, M.H., et al., *Deletion and acquisition of genomic content during early stage adaptation of Pseudomonas aeruginosa to a human host environment.* Environ Microbiol, 2012. **14**(8): p. 2200-11.

40. Marvig, R.L., Johansen, H.K., Molin, S., Jelsbak, L., *Genome Analysis of a Transmissible Lineage of Pseudomonas aeruginosa Reveals Pathoadaptive Mutations and Distinct Evolutionary Paths of Hypermutators.* PLoS Genet, 2013. **9**(9): p. e1003741.

41. Bye, M.R., J.M. Ewig, and L.M. Quittell, *Cystic fibrosis.* Lung, 1994. **172**(5): p. 251-70.

42. Riordan, J.R., et al., *Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA.* Science, 1989. **245**(4922): p. 1066-73.

43. Boucher, R.C., *Relationship of airway epithelial ion transport to chronic bronchitis.* Proc Am Thorac Soc, 2004. **1**(1): p. 66-70.

44. Klepac-Ceraj, V., et al., *Relationship between cystic fibrosis respiratory tract bacterial communities and age, genotype, antibiotics and Pseudomonas aeruginosa.* Environ Microbiol, 2010. **12**(5): p. 1293-303.

45. Rogers, G.B., et al., *characterization of bacterial community diversity in cystic fibrosis lung infections by use of 16s ribosomal DNA terminal restriction fragment length polymorphism profiling.* J Clin Microbiol, 2004. **42**(11): p. 5176-83.

46. Harris, J.K., et al., *Molecular identification of bacteria in bronchoalveolar lavage fluid from children with cystic fibrosis.* Proc Natl Acad Sci U S A, 2007. **104**(51): p. 20529-33.

47. Goddard, A.F., et al., *Direct sampling of cystic fibrosis lungs indicates that DNA-based analyses of upper-airway specimens can misrepresent lung microbiota.* Proc Natl Acad Sci U S A, 2012. **109**(34): p. 13769-74.

48. Shapiro, E.D., et al., *Bacteriology of the maxillary sinuses in patients with cystic fibrosis.* J Infect Dis, 1982. **146**(5): p. 589-93.

49. Folkesson, A., et al., *Adaptation of Pseudomonas aeruginosa to the cystic fibrosis airway: an evolutionary perspective.* Nat Rev Microbiol, 2012. **10**(12): p. 841-51.

50. Ulrich, M., et al., *Alveolar inflammation in cystic fibrosis.* J Cyst Fibros, 2010. **9**(3): p. 217-27.

51. Johansen, H.K., et al., *Colonisation and infection of the paranasal sinuses in cystic fibrosis patients is accompanied by a reduced PMN response.* J Cyst Fibros, 2012. **11**(6): p. 525-31.

52. Hull, J., et al., *Pulmonary oxidative stress response in young children with cystic fibrosis.* Thorax, 1997. **52**(6): p. 557-60.

53. Hoiby, N. and B. Frederiksen, *Microbiology of cystic fibrosis*, in *Cystic fibrosis*, M. Hodson and D. Geddes, Editors. 2000, Arnold: London, United Kingdom. p. 83-107.

54. Nguyen, D. and P.K. Singh, *Evolving stealth: genetic adaptation of Pseudomonas aeruginosa during cystic fibrosis infections.* Proc Natl Acad Sci U S A, 2006. **103**(22): p. 8305-6.

55. Doring, G., et al., *Antibiotic therapy against Pseudomonas aeruginosa in cystic fibrosis: a European consensus.* Eur Respir J, 2000. **16**(4): p. 749-67.

56. Hansen, C.R., et al., *Long-term azitromycin treatment of cystic fibrosis patients with chronic Pseudomonas aeruginosa infection; an observational cohort study.* J Cyst Fibros, 2005. **4**(1): p. 35-40.

57. Hansen, C.R., T. Pressler, and N. Hoiby, *Early aggressive eradication therapy for intermittent Pseudomonas aeruginosa airway colonization in cystic fibrosis patients: 15 years experience.* J Cyst Fibros, 2008. **7**(6): p. 523-30.

58. Hoiby, N., *Recent advances in the treatment of Pseudomonas aeruginosa infections in cystic fibrosis.* BMC Med, 2011. **9**: p. 32.

59. Geller, D.E., et al., *Pharmacokinetics and bioavailability of aerosolized tobramycin in cystic fibrosis.* Chest, 2002. **122**(1): p. 219-26.

60. Ratjen, F., et al., *Pharmacokinetics of inhaled colistin in patients with cystic fibrosis.* J Antimicrob Chemother, 2006. **57**(2): p. 306-11.

61. Aanaes, K., et al., *Decreased mucosal oxygen tension in the maxillary sinuses in patients with cystic fibrosis.* J Cyst Fibros, 2011. **10**(2): p. 114-20.

62. Worlitzsch, D., et al., *Effects of reduced mucus oxygen concentration in airway Pseudomonas infections of cystic fibrosis patients.* J Clin Invest, 2002. **109**(3): p. 317-25.

63. Palmer, K.L., S.A. Brown, and M. Whiteley, *Membrane-bound nitrate reductase is required for anaerobic growth in cystic fibrosis sputum.* J Bacteriol, 2007. **189**(12): p. 4449-55.

64. Barth, A.L. and T.L. Pitt, *The high amino-acid content of sputum from cystic fibrosis patients promotes growth of auxotrophic Pseudomonas aeruginosa.* J Med Microbiol, 1996. **45**(2): p. 110-9.

65. Ohman, D.E. and A.M. Chakrabarty, *Utilization of human respiratory secretions by mucoid Pseudomonas aeruginosa of cystic fibrosis origin.* Infect Immun, 1982. **37**(2): p. 662-9.

66. Rau, M.H., et al., *Early adaptive developments of Pseudomonas aeruginosa after the transition from life in the environment to persistent colonization in the airways of human cystic fibrosis hosts.* Environ Microbiol, 2010. **12**(6): p. 1643-58.

67. Son, M.S., et al., *In vivo evidence of Pseudomonas aeruginosa nutrient acquisition and pathogenesis in the lungs of cystic fibrosis patients.* Infect Immun, 2007. **75**(11): p. 5313-24.

68. Skaar, E.P., *The battle for iron between bacterial pathogens and their vertebrate hosts.* PLoS Pathog, 2010. **6**(8): p. e1000949.

69. Lamont, I.L., A.F. Konings, and D.W. Reid, *Iron acquisition by Pseudomonas aeruginosa in the lungs of patients with cystic fibrosis.* Biometals, 2009. **22**(1): p. 53-60.

70. Hunter, R.C., et al., *Ferrous iron is a significant component of bioavailable iron in cystic fibrosis airways.* MBio, 2013. **4**(4).

71. Konings, A.F., et al., *Pseudomonas aeruginosa uses multiple pathways to acquire iron during chronic infection in cystic fibrosis lungs.* Infect Immun, 2013. **81**(8): p. 2697-704.

72. Gilljam, H., A. Ellin, and B. Strandvik, *Increased bronchial chloride concentration in cystic fibrosis.* Scand J Clin Lab Invest, 1989. **49**(2): p. 121-4.

73. Joris, L., I. Dab, and P.M. Quinton, *Elemental composition of human airway surface fluid in healthy and diseased airways.* Am Rev Respir Dis, 1993. **148**(6 Pt 1): p. 1633-7.

74. Damkiaer, S., et al., *Evolutionary remodeling of global regulatory networks during long-term bacterial adaptation to human hosts.* Proc Natl Acad Sci U S A, 2013. **110**(19): p. 7766-71.

75. Lyczak, J.B., C.L. Cannon, and G.B. Pier, *Establishment of Pseudomonas aeruginosa infection: lessons from a versatile opportunist.* Microbes Infect, 2000. **2**(9): p. 1051-60.

76. Stover, C.K., et al., *Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen.* Nature, 2000. **406**(6799): p. 959-64.

77. Winsor, G.L., et al., *Pseudomonas Genome Database: improved comparative analysis and population genomics capability for Pseudomonas genomes.* Nucleic Acids Res, 2011. **39**(Database issue): p. D596-600.

78. Wiehlmann, L., et al., *Population structure of Pseudomonas aeruginosa.* Proc Natl Acad Sci U S A, 2007. **104**(19): p. 8101-6.

79. Hansen, S.K., et al., *Evolution and diversification of Pseudomonas aeruginosa in the paranasal sinuses of cystic fibrosis children have implications for chronic lung infection.* ISME J, 2012. **6**(1): p. 31-45.

80. Jelsbak, L., et al., *Molecular epidemiology and dynamics of Pseudomonas aeruginosa populations in lungs of cystic fibrosis patients.* Infect Immun, 2007. **75**(5): p. 2214-24.

81. Cramer, N., et al., *Microevolution of the major common Pseudomonas aeruginosa clones C and PA14 in cystic fibrosis lungs.* Environ Microbiol, 2011. **13**(7): p. 1690-704.

82. Klockgether, J., et al., *Intraclonal diversity of the Pseudomonas aeruginosa cystic fibrosis airway isolates TBCF10839 and TBCF121838: distinct signatures of transcriptome, proteome, metabolome, adherence and*

pathogenicity despite an almost identical genome sequence. Environ Microbiol, 2012.

83. Burns, J.L., et al., *Longitudinal assessment of Pseudomonas aeruginosa in young children with cystic fibrosis.* J Infect Dis, 2001. **183**(3): p. 444-52.

84. Mahenthiralingam, E., M.E. Campbell, and D.P. Speert, *Nonmotility and phagocytic resistance of Pseudomonas aeruginosa isolates from chronically colonized patients with cystic fibrosis.* Infect Immun, 1994. **62**(2): p. 596-605.

85. Godfrey, A.J., L.E. Bryan, and H.R. Rabin, *beta-Lactam-resistant Pseudomonas aeruginosa with modified penicillin-binding proteins emerging during cystic fibrosis treatment.* Antimicrob Agents Chemother, 1981. **19**(5): p. 705-11.

86. Taylor, R.F., M.E. Hodson, and T.L. Pitt, *Adult cystic fibrosis: association of acute pulmonary exacerbations and increasing severity of lung disease with auxotrophic mutants of Pseudomonas aeruginosa.* Thorax, 1993. **48**(10): p. 1002-5.

87. Doggett, R.G., G.M. Harrison, and E.S. Wallis, *Comparison of Some Properties of Pseudomonas Aeruginosa Isolated from Infections in Persons with and without Cystic Fibrosis.* J Bacteriol, 1964. **87**: p. 427-31.

88. Oliver, A., et al., *High frequency of hypermutable Pseudomonas aeruginosa in cystic fibrosis lung infection.* Science, 2000. **288**(5469): p. 1251-4.

89. Luzar, M.A. and T.C. Montie, *Avirulence and altered physiological properties of cystic fibrosis strains of Pseudomonas aeruginosa.* Infect Immun, 1985. **50**(2): p. 572-6.

90. Ernst, R.K., et al., *Specific lipopolysaccharide found in cystic fibrosis airway Pseudomonas aeruginosa.* Science, 1999. **286**(5444): p. 1561-5.

91. Cabral, D.A., B.A. Loh, and D.P. Speert, *Mucoid Pseudomonas aeruginosa resists nonopsonic phagocytosis by human neutrophils and macrophages.* Pediatr Res, 1987. **22**(4): p. 429-31.

92. Meshulam, T., et al., *Phagocytosis of mucoid and nonmucoid strains of Pseudomonas aeruginosa.* Clin Immunol Immunopathol, 1984. **32**(2): p. 151-65.

93. Govan, J.R., *Antibiotic therapy and cystic fibrosis: increased resistance of mucoid Pseudomonas aeruginosa to carbenicillin.* J Antimicrob Chemother, 1976. **2**(2): p. 215-7.

94. Burns, M.W. and J.R. May, *Bacterial precipitins in serum of patients with cystic fibrosis.* Lancet, 1968. **1**(7537): p. 270-2.

95. Martin, D.W., et al., *Mechanism of conversion to mucoidy in Pseudomonas aeruginosa infecting cystic fibrosis patients.* Proc Natl Acad Sci U S A, 1993. **90**(18): p. 8377-81.

96. Breidenstein, E.B., C. de la Fuente-Nunez, and R.E. Hancock, *Pseudomonas aeruginosa: all roads lead to resistance.* Trends Microbiol, 2011. **19**(8): p. 419-26.

97. Pai, H., et al., *Carbapenem resistance mechanisms in Pseudomonas aeruginosa clinical isolates.* Antimicrob Agents Chemother, 2001. **45**(2): p. 480-4.

98. Llanes, C., et al., *Clinical strains of Pseudomonas aeruginosa overproducing MexAB-OprM and MexXY efflux pumps simultaneously.* Antimicrob Agents Chemother, 2004. **48**(5): p. 1797-802.

99. Ballestero, S., et al., *Carbapenem resistance in Pseudomonas aeruginosa from cystic fibrosis patients.* J Antimicrob Chemother, 1996. **38**(1): p. 39-45.

100. Schurek, K.N., et al., *Involvement of pmrAB and phoPQ in polymyxin B adaptation and inducible resistance in non-cystic fibrosis clinical isolates of Pseudomonas aeruginosa.* Antimicrob Agents Chemother, 2009. **53**(10): p. 4345-51.

101. Barrow, K. and D.H. Kwon, *Alterations in two-component regulatory systems of phoPQ and pmrAB are associated with polymyxin B resistance in clinical isolates of Pseudomonas aeruginosa.* Antimicrob Agents Chemother, 2009. **53**(12): p. 5150-4.

102. Abraham, N. and D.H. Kwon, *A single amino acid substitution in PmrB is associated with polymyxin B resistance in clinical isolate of Pseudomonas aeruginosa.* FEMS Microbiol Lett, 2009. **298**(2): p. 249-54.

103. Pasca, M.R., et al., *Evaluation of fluoroquinolone resistance mechanisms in Pseudomonas aeruginosa multidrug resistance clinical isolates.* Microb Drug Resist, 2012. **18**(1): p. 23-32.

104. Cabot, G., et al., *Genetic markers of widespread extensively drug-resistant Pseudomonas aeruginosa high-risk clones.* Antimicrob Agents Chemother, 2012. **56**(12): p. 6349-57.

105. Juan, C., et al., *Molecular mechanisms of beta-lactam resistance mediated by AmpC hyperproduction in Pseudomonas aeruginosa clinical strains.* Antimicrob Agents Chemother, 2005. **49**(11): p. 4733-8.

106. De Vos, D., et al., *Study of pyoverdine type and production by Pseudomonas aeruginosa isolated from cystic fibrosis patients: prevalence of type II pyoverdine isolates and accumulation of pyoverdine-negative mutations.* Arch Microbiol, 2001. **175**(5): p. 384-8.

107. Grimwood, K., et al., *Elevated exoenzyme expression by Pseudomonas aeruginosa is correlated with exacerbations of lung disease in cystic fibrosis.* Pediatr Pulmonol, 1993. **15**(3): p. 135-9.

108. Mena, A., et al., *Genetic adaptation of Pseudomonas aeruginosa to the airways of cystic fibrosis patients is catalyzed by hypermutation.* J Bacteriol, 2008. **190**(24): p. 7910-7.

109. Waine, D.J., et al., *Association between hypermutator phenotype, clinical variables, mucoid phenotype, and antimicrobial resistance in Pseudomonas aeruginosa.* J Clin Microbiol, 2008. **46**(10): p. 3491-3.

110. Ferroni, A., et al., *Effect of mutator P. aeruginosa on antibiotic resistance acquisition and respiratory function in cystic fibrosis.* Pediatr Pulmonol, 2009. **44**(8): p. 820-5.

111. Ciofu, O., et al., *Occurrence of hypermutable Pseudomonas aeruginosa in cystic fibrosis patients is associated with the oxidative stress caused by chronic lung inflammation.* Antimicrob Agents Chemother, 2005. **49**(6): p. 2276-82.

112. Moyano, A.J., et al., *MutS deficiency and activity of the error-prone DNA polymerase IV are crucial for determining mucA as the main target for mucoid conversion in Pseudomonas aeruginosa.* Mol Microbiol, 2007. **64**(2): p. 547-59.

113. Elena, S.F. and R.E. Lenski, *Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation.* Nat Rev Genet, 2003. **4**(6): p. 457-69.

114. Langergraber, K.E., et al., *Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution.* Proc Natl Acad Sci U S A, 2012. **109**(39): p. 15716-21.

115. Morales, G., et al., *Structure of Pseudomonas aeruginosa populations analyzed by single nucleotide polymorphism and pulsed-field gel electrophoresis genotyping.* J Bacteriol, 2004. **186**(13): p. 4228-37.

116. Marvig, R.L., et al., *Mutations in 23S rRNA confer resistance against azithromycin in Pseudomonas aeruginosa.* Antimicrob Agents Chemother, 2012.

117. Gaffe, J., et al., *Insertion sequence-driven evolution of Escherichia coli in chemostats.* J Mol Evol, 2011. **72**(4): p. 398-412.

118. Marvig, R.L., Damkiær, S., Khademi, H., Markussen, T., Molin, S., Jelsbak, L., *Within-Host Evolution of Pseudomonas aeruginosa Reveals Adaptation Towards Iron Acquisition from Hemoglobin.* Unpublished, 2013.

Chapter 6

Research papers

The research articles that are included in this thesis are enclosed in the following order.

Article 1

Yang L*, Jelsbak L*, **Marvig RL**, Damkiær S, Workman CT, Rau MH, Hansen SK, Folkesson A, Johansen HK, Ciofu O, Høiby N, Sommer MO, Molin S. Evolutionary dynamics of bacteria in a human host environment. *Proc Natl Acad Sci U S A*. 2011 May 3;108(18):7481-6.

Article 2

Marvig RL*, Søndergaard MS*, Damkiær S, Høiby N, Johansen HK, Molin S, Jelsbak L. Mutations in 23S rRNA confer resistance against azithromycin in *Pseudomonas aeruginosa*. *Antimicrob Agents Chemother*. 2012. 56(8):4519-21.

Article 3

Rau MH, **Marvig RL***, Ehrlich GD, Molin S, Jelsbak L. Deletion and acquisition of genomic content during early stage adaptation of *Pseudomonas aeruginosa* to a human host environment. *Environ Microbiol*. 2012 Aug;14(8):2200-11.

Article 4

Marvig RL, Johansen HK, Molin S, Jelsbak L. Genome Analysis of a Transmissible Lineage of *Pseudomonas aeruginosa* Reveals Pathoadaptive Mutations and Distinct Evolutionary Paths of Hypermutators. *PLoS Genet*. 2013. 9(9):e1003741.

Article 5

Marvig RL*, Damkiær S*, Khademi H, Markussen, Molin S, Jelsbak L. Within-Host Evolution of Pseudomonas aeruginosa Reveals Adaptation Towards Iron Acquisition from Hemoglobin. *Unpublished*. 2013.

* Denotes equal contribution.

Evolutionary dynamics of bacteria in a human host environment

Lei Yang[a,1], Lars Jelsbak[a,1], Rasmus Lykke Marvig[a], Søren Damkiær[a], Christopher T. Workman[a], Martin Holm Rau[a], Susse Kirkelund Hansen[a], Anders Folkesson[a], Helle Krogh Johansen[b], Oana Ciofu[c], Niels Høiby[b,c], Morten O. A. Sommer[a,d], and Søren Molin[a,2]

[a]Department of Systems Biology, Technical University of Denmark, 2800 Lyngby, Denmark; [b]Department of Clinical Microbiology, Rigshospitalet, University Hospital, 2100 Copenhagen, Denmark; [c]Institute for International Health, Immunology and Microbiology, University of Copenhagen, 2200 Copenhagen, Denmark; and [d]Department of Genetics, Harvard Medical School, Boston, MA 02115

Edited* by Richard E. Lenski, Michigan State University, East Lansing, MI, and approved March 25, 2011 (received for review December 7, 2010)

Laboratory evolution experiments have led to important findings relating organism adaptation and genomic evolution. However, continuous monitoring of long-term evolution has been lacking for natural systems, limiting our understanding of these processes in situ. Here we characterize the evolutionary dynamics of a lineage of a clinically important opportunistic bacterial pathogen, *Pseudomonas aeruginosa*, as it adapts to the airways of several individual cystic fibrosis patients over 200,000 bacterial generations, and provide estimates of mutation rates of bacteria in a natural environment. In contrast to predictions based on in vitro evolution experiments, we document limited diversification of the evolving lineage despite a highly structured and complex host environment. Notably, the lineage went through an initial period of rapid adaptation caused by a small number of mutations with pleiotropic effects, followed by a period of genetic drift with limited phenotypic change and a genomic signature of negative selection, suggesting that the evolving lineage has reached a major adaptive peak in the fitness landscape. This contrasts with previous findings of continued positive selection from long-term in vitro evolution experiments. The evolved phenotype of the infecting bacteria further suggests that the opportunistic pathogen has transitioned to become a primary pathogen for cystic fibrosis patients.

microbial evolution | natural population | chronic infection | genome sequences | transcriptomics

Bacterial populations maintained in defined in vitro systems are well-suited for experimental evolutionary studies of adaptation to novel environments. It has been observed from such investigations that the initial rate of adaptation is rapid with a continuous increase in fitness that later tends to decelerate over time (1). Experimental evolution has also been used to gain insight into the molecular details of evolution (2–7). Interestingly, a complex relationship between the decelerating rate of evolutionary adaptation and a constant rate of genomic changes enriched with nonsynonymous changes has been observed in one such in vitro experiment (2). The level of diversity that evolves in a population is thought to increase as a function of environmental complexity in terms of available ecological niches. Laboratory experiments with artificially structured microbial populations have shown that environmental heterogeneity, such as population complexity and spatial structure, can generate and maintain population diversity (8–10). Comparable systematic investigations of bacterial populations living in complex natural environments are lacking, and the extent to which laboratory evolutionary experiments bear any relevance to evolutionary processes in natural environments is poorly understood. To address this question, we investigated the evolutionary dynamics including the roles of neutral versus selective mechanisms in bacterial populations evolving over 200,000 generations in the complex natural and structured environment of cystic fibrosis (CF) airways.

Chronic airway infections in CF patients by the bacterium *Pseudomonas aeruginosa* offer a rare opportunity to study evolutionary adaptation of bacteria in natural environments. CF airways, like other natural environments, constitute a complex and highly dynamic ecosystem composed of a range of niches that vary in space and time. Factors such as heterogeneous distribution of oxygen, nutrients, and other substances, fluctuations in exposure to the host immune system and therapeutic antibiotic treatments, as well as a diverse microbial community structure in the host all contribute to environmental conditions under which the bacteria face continuous adaptive challenges (11). These environmental conditions have been used to explain the extensive phenotypic diversity often observed in *P. aeruginosa* populations from the same CF respiratory specimen (12–14). We propose that *P. aeruginosa* infections in CF airways constitute a valuable natural system for addressing central questions concerning evolutionary biology, and we use this system to investigate how the highly successful and transmissible DK2 lineage of *P. aeruginosa* adapts to this environmental complexity in several CF patients using genome sequencing, transcriptional profiling, and phenotypic arrays. Based on previous estimates of in situ growth rates of *P. aeruginosa* isolates (15), we calculate that the DK2 lineage has persisted for about 200,000 bacterial generations in the CF airways of several patients, and as such this study documents the genomic and functional details of a long microbial evolution "experiment."

Results and Discussion

This study was based on a collection of *P. aeruginosa* isolates sampled from several hundred Danish CF patients between 1973 and 2008. Initial examination of the clonal relationship among a subset of the stored isolates using molecular typing methods revealed that some *P. aeruginosa* clonal lineages were observed repeatedly because of transmission among the patients (16). For the present study, we are focusing on the DK2 lineage (*Materials and Methods*), which we have established as one of these transmissible lineages and isolated from ~40 patients since the start of the sampling program in 1973.

Genomic Evolution of the DK2 Lineage. To enable a detailed characterization of the evolutionary trajectory associated with the dissemination of the DK2 lineage across multiple patients, we sequenced the genomes of 12 selected DK2 strains (CF114-1973,

Author contributions: L.J. and S.M. conceived study; L.Y., L.J., M.O.A.S., and S.M. designed research; L.Y., L.J., R.L.M., S.D., M.H.R., and M.O.A.S. performed research; H.K.J., O.C., and N.H. contributed new reagents/analytic tools; L.Y., L.J., R.L.M., S.D., C.T.W., M.H.R., S.K.H., A.F., M.O.A.S., and S.M. analyzed data; and L.Y., L.J., R.L.M., S.D., A.F., M.O.A.S., and S.M. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]L.Y. and L.J. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: sm@bio.dtu.dk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1018249108/-/DCSupplemental.

EVOLUTION

CF43-1973, CF66-1973, CF66-1992, CF66-2008, CF30-1979, CF173-1984, CF173-2005, CF333-1991, CF333-1997, CF333-2003, and CF333-2007) isolated from 6 different patients between 1973 and 2008 (Fig. 1 and Tables S1 and S2). We identified high-quality SNPs in the nonrepetitive parts of the genomes by mapping sequence reads for each isolate against CF333-2007 (Table S3 and *SI Materials and Methods*). A total of 368 SNPs have appeared since the clones diverged from their common ancestor before 1973 (Table S3), and we find that there are 180 SNPs between the most distantly related clones (CF114-1973 and CF333-2007). Based on the identified SNPs, we construct a tree representing the evolutionary relationship of the DK2 strains (Fig. 2*A* and *Materials and Methods*), and as such the tree demonstrates the sequence of mutational events in the DK2 lineage from 1973 and onward.

We predict the DK2 ancestor to be related to the sequenced strains, as indicated with the predicted root in Fig. 2*A*, based on the assumption that the DK2 ancestor would contain wild-type SNP alleles [i.e., SNP positions with sequences identical to homolog sequences found in the *P. aeruginosa* reference strains PAO1 and PA14 (17)]. Despite the genomic diversity among the 1973 isolates, their phenotypes are similar to the well-characterized laboratory *P. aeruginosa* strain PAO1 (Fig. 3 and Table S1). Because the CF infection process is associated with the appearance of phenotypes of which many are not usually observed among environmental, wild-type isolates of this species (18–20), we predict that the immediate ancestor to the DK2 lineage entered the CF environment shortly before the first sampling in 1973. Because the original infecting DK2 strain is unavailable for analysis, we cannot precisely determine the pre-1973 events that resulted in genomic diversity among the isolates sampled in 1973. Although a complete understanding of the pre-1973 history of the DK2 lineage is not a prerequisite for the present study, we hypothesize that this diversity may be a result of processes such as (*i*) an already-present diversity in the ancestral DK2 lineage before first entry into the CF environment (i.e., the first infection caused by a mix of genetically distinct subpopulations of DK2) and/or a rapid diversification of the ancestral DK2 strain upon first infection as observed in patients infected with environmental strains (13, 19) possibly coupled to a transmission process that retained diversity, (*ii*) the storage conditions of these particular isolates in agar slants at room temperature for an extended period in which genomic changes may have arisen (21), or (*iii*) an increased mutation rate in the lineage before the 1973 sampling. The presence of a missense mutation in *mutY* and a molecular signature consistent with a MutY defect [i.e., a high proportion of transversions ($P = 7 \times 10^{-13}$) of which 85% are G·C→T·A] in the sublineage branching into CF43-1973 indeed demonstrates a pre-1973 hypermutator phenotype for this particular isolate (22). As the two remaining 1973 strains did not show molecular signs of a hypermutator phenotype, other phenomena may have led to increased genomic diversification in these strains before sampling in 1973.

We derive two conclusions based on the topology of the tree. First, all isolates sampled from 1979 and onward share a common branch point (marked with a red circle in Fig. 2*A*). We hypothesize that important mutations accumulated in the early hosts (pre-1979) positioned the DK2 lineage for subsequent reproductive success in the CF airways in several hosts. Second, the tree demonstrates a remarkably limited diversification of the sequential isolates present in patients CF333 and CF66 after 1979. This observation is surprising, because CF airways represent a dynamic and spatially structured environment in which one would expect a divergent population structure and multiple coexisting sublineages resulting in a deeply branched phylogenetic tree of clones isolated from the same host (8, 12, 13, 23). Note, however, the disconnected branching of the 1984 and 2005 isolates in patient CF173 (Fig. 2*A*), which suggests that a replacement of the DK2 lineage caused by a secondary transmission event after 1984 has taken place.

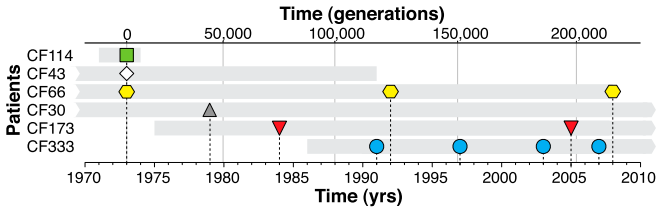


Fig. 1. Isolate sampling points and patient life span. DK2 *P. aeruginosa* isolates were collected from six different CF patients during a 35-y time period. Bacterial isolates are represented by the colored symbols, and gray bars represent patient life span.

Because the DK2 lineage has exhibited a highly successful phenotype over an estimated period of 200,000 generations of growth, it provides an excellent system for the detailed study of molecular evolution in ways that previously have been possible only in well-defined artificial laboratory environments. Importantly, we find that SNPs in the protein-coding regions of the genome have accumulated at a constant rate in the DK2 lineages since 1973 (Fig. 2 *B* and *C*). We can estimate the mutation rate based on the accumulation of synonymous mutations in the protein-coding regions of the genome assuming that they are neutral and not selected for (2, 24). Based on the codon use in *P. aeruginosa* (25), we find that 25% of all SNPs in the protein-coding parts of the genome result in synonymous changes (*SI Materials and Methods*). We fit a least-squares linear model of mutation accumulation over time and calculate the mutation rate as the number of synonymous mutations fixed per generation (cf. Fig. 2*B*) divided by the fraction of mutations expected to be synonymous in the genome. Based on these assumptions, we find that the mutation rate of *P. aeruginosa* in CF airways is 7.2×10^{-11} SNPs/bp per generation, which is in the range of previous estimates for related bacteria (2, 26, 27). Excluding the 1973 isolates from the calculation, we find the mutation rate to be 6.9×10^{-11} SNPs/bp per generation. It is important to note that these estimates show no indication of increased rates of mutation, as has been suggested for biofilm-associated cells and CF infection-associated cells due to oxidative stress or other phenomena (28–32).

Selective Forces Acting on the DK2 Genome. Based on the number of protein-coding synonymous SNPs, we can estimate the expected number of nonsynonymous protein-coding SNPs assuming a neutral theory of molecular evolution. This theory predicts the rate of nonsynonymous substitutions (dN) relative to the rate of synonymous substitutions (dS) to approach unity during neutral evolution. Over- or underrepresentation of nonsynonymous mutations can be seen as a sign of positive or negative selection, respectively. Overall, among all mutations in the protein-coding regions of the genome, we find a significant underrepresentation of nonsynonymous mutations (dN/dS = 0.79; $P = 1.3 \times 10^{-4}$) (Table S4). The sign of negative selection is most prevalent in isolates from 1979 and onward (branches G-U in Fig. 2*A*; dN/dS = 0.70; $P = 6.5 \times 10^{-5}$), whereas calculations based on early isolates (branches A–F in Fig. 2*A*) do not support evidence of negative selection (dN/dS = 0.89; $P = 0.10$). This observation supports a hypothesis that a selection of adaptive mutations takes place in the years up to 1979 leading to a highly adapted clone closely related to CF30-1979 (represented by a red circle in Fig. 2*A*). Although our results contrast with previous findings of strong positive selection in CF airways (23), it is likely that the earliest stages (before 1973) of infection by the DK2 lineage were also dominated by positive selection mechanisms. Indeed, despite of the overall sign of negative selection, there is a statistically significant sign of positive selection in the lineage (branch F in Fig. 2*A*) leading to the post-1979 isolates (dN/dS = 1.58; $P = 0.02$).

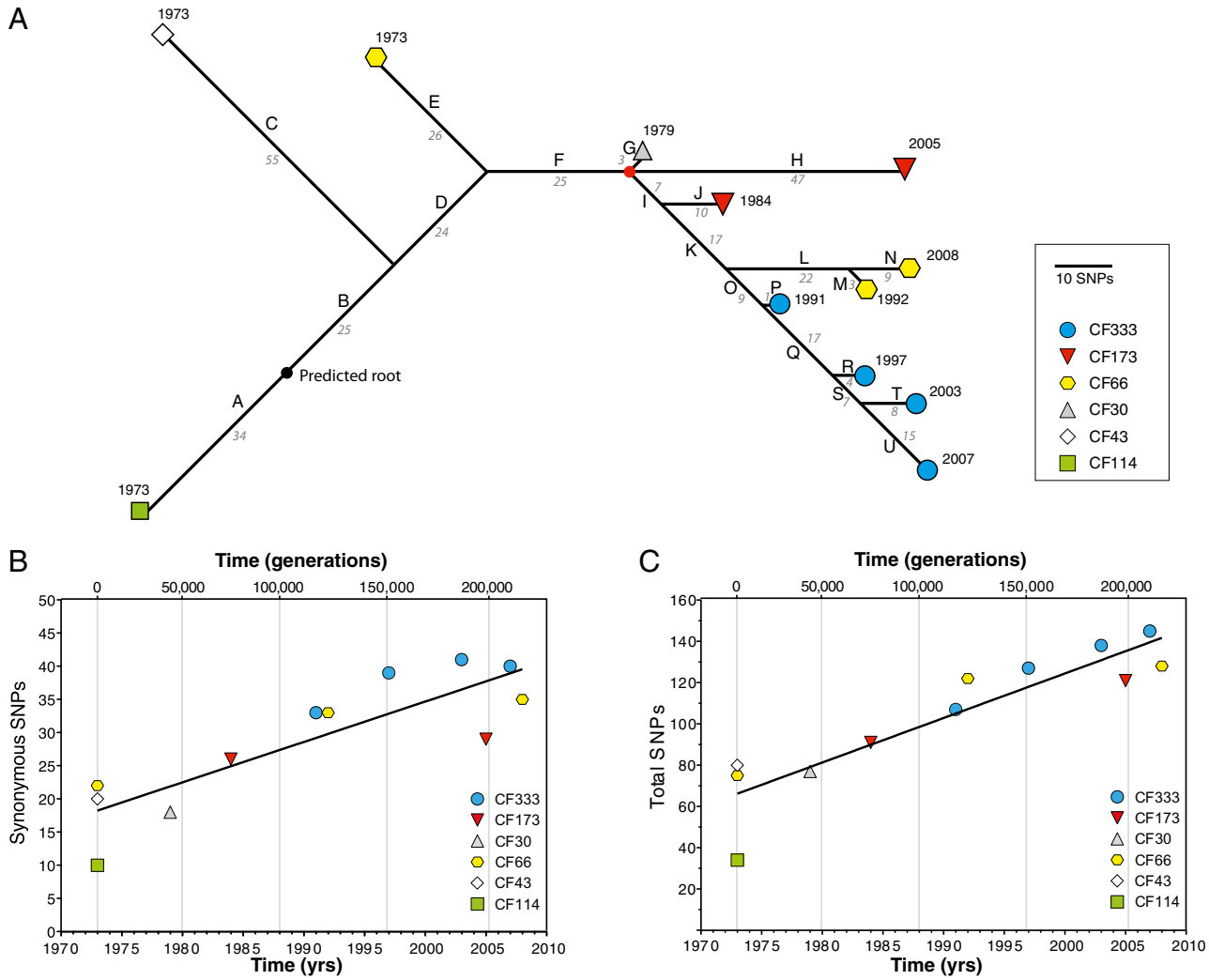


Fig. 2. Evolutionary trajectory of the DK2 lineage. (*A*) Tree representing the evolutionary relationship of the DK2 clones. The tree is based on the accumulation of SNPs identified from genome sequencing. Lengths of branches are proportional to the SNPs between the isolates as indicated by the numbers in italics. The black circle that separates branches "A" and "B" represents the predicted root (i.e., the immediate common ancestor of the isolates). Synonymous (*B*) and total (*C*) SNPs accumulated between the predicted root and the tip of each branch are plotted against time (and estimated generations).

Phenotypic Changes During Long-Term Adaptation to CF Airways. The dynamics of phenotype changes in the isolates of the DK2 lineage was characterized both by gene expression profiles under a single growth condition (standard laboratory medium) and by determination of catabolic activities under different growth conditions. Each method interrogates different aspects of cellular physiology (whole-genome transcript abundance and cellular catabolic capacity) and is therefore expected to probe different components of the genotype/phenotype relationship of the bacteria.

Hierarchical clustering of all recorded expression and catabolic signals from each isolate showed that strains sampled between 1979 and 2008 are more similar to each other than to the initial isolates (1973) (Fig. S1 *A* and *B*). To further examine the extent of gene expression changes among the isolates a principal component analysis, which transforms a complex data matrix into a small number of variables, was applied. The first principal component PC1 accounts for most (42%) of the variance in the dataset. Plotting of the loadings of each isolate on PC1 as a function of time shows a strikingly simple and clear time-dependent progression, with a major shift between 1973 and 1979 followed by a period with few changes (Fig. 3*A*). After 1979 all of the DK2 isolates share a very similar gene expression pattern despite the fact that they cover samples from four different patients taken with a time span of 29 y (estimated to correspond to ~150,000 bacterial generations) and have accumulated 179 mutations.

In analogy with the global transcriptional data, the "catabolome" datasets display a first principal component diagram that documents a lack of variation among strains isolated after 1979 (Fig. 3*B*). Before 1979 the profile changed rapidly from a wild-type pattern of catabolism of C and N sources, as observed in a reference strain (PAO1) and in the 1973 isolates, to the adapted pattern reflecting a mutation(s) in a global regulatory gene(s). Surprisingly, the catabolome as monitored by the Biolog phenotype arrays shows only loss of function (Fig. S2), which is paradoxical, taking the abundance of, for example, amino acids in the sputum into consideration (33). We hypothesize that the airway mucus and the surrounding environment is rich in a variety of nutrients and that the competition for these from other infecting species is marginal, providing ample opportunities for alternative feeding by *P. aeruginosa*. Overall, our phenotype data are consistent with our genomic data and suggest strong negative selective forces acting on the evolving populations of the DK2 lineage after 1979.

The observed phenotypic stability and the genomic signature of negative selection are consistent with a hypothesis that the

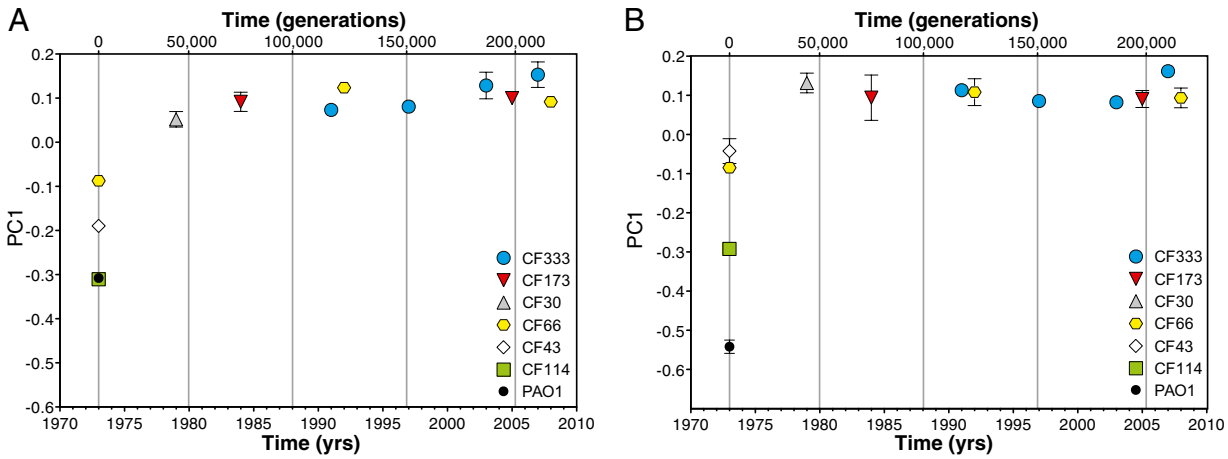


Fig. 3. Phenotypic relationships and dynamics among DK2 isolates. Principal component analyses of the microarray expression data (*A*) and Biolog phenotypic data (*B*) show a similar pattern suggesting that major changes occurred before 1979. First principal component (PC1) loadings for each isolate are graphed as a function of time (and estimated generations). PC1 represents 42% (*A*) and 63% (*B*) of the total variation among isolates. Errors bars indicate standard deviations of biological triplicates (*A*) or duplicates (*B*) of the same isolate.

DK2 lineage (represented by the evolved genotype/phenotype of CF30-1979 and later isolates) becomes positioned on a major fitness peak in this complex and dynamic natural environment. If the evolved DK2 lineage is in fact positioned on a major fitness peak, most new mutations would result in reduced fitness and therefore not be fixed in the population (negative selection), whereas neutral (synonymous) mutations would be fixed (genetic drift), which is in agreement with our findings.

Parallel Evolution in DK2 Sublineages. The striking similarity between the phenotypes of DK2 isolates sampled from different CF patients suggests that the host environments represent parallel selective conditions by which evolution is directed. Such parallelism is expected to generate parallel evolution (20), and a search for signs of independent evolution of the same traits in the isolates confirms this. In the case of two phenotypes that evolved independently in the CF173 and CF333 isolates after their early separation from CF30-1979, we were able to identify the genetic details: (*i*) The lost capacity of both isolates to catabolize 4-hydroxyphenylacetic acid (4-HPA), a trait identified in the Biolog phenotypic arrays, is associated with similar—but not identical—genetic changes (Fig. 4 *A*–*C*); and (*ii*) increased resistance to the antibiotic ciprofloxacin conferred by two independent *gyrA* mutations that have previously been described to contribute to ciprofloxacin resistance (34) (Fig. 4*D*). Consistent with this, both isolates CF173-2005 and CF333-2007 show increased levels of resistance toward ciprofloxacin (Fig. 4*D*). Additionally, both isolates obtained mutations in the genes PA5160 and *mexY*, which are components of multidrug transport systems (35) (Fig. 4*A*).

Mutations with Pleiotropic Effects Drive Adaptation to New Environments and Reproductive Success in CF Airways. The genomic analysis showed that mutations appeared in several global regulatory genes during the initial 42,000 estimated generations of growth. Indeed, we observed more missense and nonsense SNPs in regulatory genes in the first 42,000 estimated generations before 1979 (11 mutations) compared with the following 167,000 estimated generations after 1979 (5 mutations) (Table S5). Of particular interest are mutations in the *mucA* (*SI Materials and Methods*) and *rpoN* genes that occurred before 1973 (as observed in CF66-1973), and a deletion of the *lasR* gene that occurred before 1979 (as observed in CF30-1979). Gene expression controlled by MucA, LasR, and RpoN accounts for approximately half of the total significantly changed gene expression between CF114-1973 and CF30-1979, which confirms the pleiotropy of these mutations in the DK2 lineage (*SI Materials and Methods*). Notably, transcomplementation of late DK2 isolates with wild-type RpoN expressed from a recombinant plasmid restored to a large extent the catabolic profile of wild-type *P. aeruginosa* (Fig. S2), suggesting that much of the adapted catabolome activity in DK2 is directly associated with the *rpoN* mutation already present in the CF66-1973 isolate. After 1979 no similarly dramatic global regulatory changes seem to have occurred in the DK2 lineage.

Based on the phenotypic data described above and tracking of the mutated alleles of *mucA*, *lasR*, and *rpoN* among DK2 isolates, we conclude that all DK2 clones isolated after 1979 from more than 40 different patients have a common ancestor with a genotype and phenotype like the one described for the CF30-1979 isolate (*SI Materials and Methods*). These findings show that extensive transmission of the DK2 lineage has occurred in the

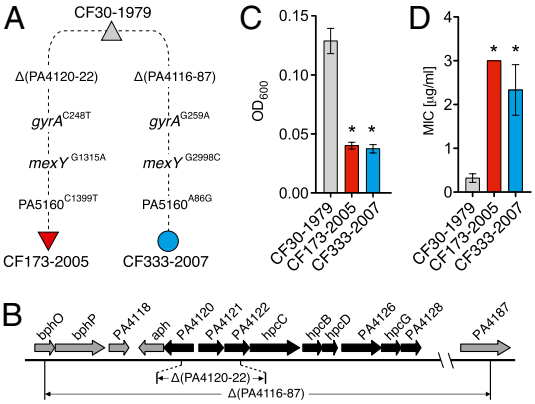


Fig. 4. Parallel evolution in CF333 and CF173 isolates. (*A*) Schematic representation of the four parallel, independent genetic changes observed in the DK2 lineage in CF333 and CF173. The genetic changes include missense/nonsense mutations in three genes (*gyrA*, *mexY*, and PA5160) and deletions of genes involved in 4-HPA catabolism. (*B*) Outline of the chromosomal region with genes required for 4-HPA catabolism (labeled in black) and positions of the deletions Δ(PA4116-87) (CF333) and Δ(PA4120-22) (CF173). (*C*) Growth of strains in minimal medium with 4-HPA as carbon source. (*D*) Ciprofloxacin minimal inhibitory concentration (MIC; μg/mL) values measured using Etest strips (AB Biodisk). Error bars represent standard deviations of three replicate experiments. Asterisks show significant differences ($P < 0.05$).

clinic after the occurrence of the last (*lasR*) of the three global regulatory mutations. Interestingly, in CF66, in whom the original DK2 strain found in 1973 went extinct (either by successful treatment or by competition with other genotypes of *P. aeruginosa* or both), a later transmission of an adapted DK2 lineage (the CF30-1979 adapted phenotype) occurred in 1992. This further supports that reproductive success (fitness) of the DK2 lineage was dependent on the particular "genetic configuration" found in CF30-1979, and that components of this fitness can be described by the phenotypes measured here.

The infecting bacteria in CF airways represent a continuous challenge to the immune system, and it is striking that among the many transcriptional changes imposed by the three global regulatory mutations (*mucA*, *rpoN*, *lasR*), reduced gene expression in the DK2 isolates is in fact associated with *P. aeruginosa* antigenic virulence factors (19, 36, 37). It is well-known that *lasR* and *rpoN* regulate virulence in this organism, and recently we documented that also *mucA* exerts control of expression of several virulence genes (19). It is therefore tempting to speculate that the combined reduction of virulence caused by the three early regulatory mutations is a key factor in the success of the DK2 lineage of *P. aeruginosa*. It should also be noted that very few regulatory mutations are identified in the lineage after 1979 (Table S5).

Concluding Remarks

We have presented a detailed analysis of the temporal sequence of all of the genetic and phenotypic changes that lead to the successful establishment of a bacterial lineage in a large number of new parallel environments. The resulting map of evolutionary changes shows that the DK2 lineage underwent an initial period of rapid adaptation upon colonization of the cystic fibrosis airways followed by a long period with limited phenotypic change and a genomic signature of negative selection.

This study of bacterial evolution in a natural environment challenges the generality of two major findings from laboratory evolution experiments. First, our data show that successful and adapted bacterial populations in natural systems can have genomic signatures of negative selection, in contrast with previous findings of continuous positive selection from evolving populations in constant, nonstructured in vitro environments (2). We note, however, that the earliest stages after the DK2 lineage migrated from the environment to the CF airways (before 1973) may have been dominated by positive selection. Second, our data demonstrate that a complex and variable environment can be dominated by a highly homogeneous population, in contrast to previous findings of diversity promotion and maintenance from laboratory experiments (8–10). These findings underscore the importance of studying the molecular details of evolving populations in natural environments in addition to populations in laboratory systems to further our understanding of basic evolutionary processes.

P. aeruginosa is an opportunistic pathogen capable of switching lifestyle as it moves from the environment to compromised hosts. The successful adaptation of the DK2 lineage in a group of CF patients in the Copenhagen CF Clinic is associated with severely reduced growth rate, loss of substantial catabolic activities and motility, and inactivation of important regulatory functions (Table S1). It is unlikely that these changes would not affect the organism's fitness in its original environment. In CF airways, however, the DK2 lineage has proven to be persistent, dominant, and transmissible. The overall adapted phenotype of DK2 is therefore one of a true pathogen, and its evolutionary history illustrates the process of how an opportunistic pathogen with a broad environmental range can transition to a primary-host-specific pathogen for CF patients. With the presented map of evolutionary changes behind this shift, it may now become possible to design novel therapeutic approaches that specifically target the adaptive process and in that way block development of chronic CF lung infection.

Materials and Methods

Bacterial Isolation. *P. aeruginosa* isolates sampled from several hundred Danish CF patients since 1973 have been stored in the Copenhagen Cystic Fibrosis Clinic. Isolation and identification of *P. aeruginosa* from sputum was done as previously described (16). The number of *P. aeruginosa* cfu in such respiratory secretions may be as high as 10^7–10^9 per mL (38). We confirmed that all strains analyzed in the present study belong to the same lineage DK2 [formerly the "b" lineage (16)] using three independent methods for determination of clonal relationships: pulsed-field gel electrophoresis, SNP typing using an ArrayTube microarray (39), and multilocus sequence typing. Single colonies were cultivated and stored at −80 °C with glycerol. The initial storage procedure for the earliest clinical isolates (1973) involved storage in agar slants at room temperature for an extended period before storage at −80 °C.

Genome Sequencing. Genomic DNA was prepared from *P. aeruginosa* isolates using a Wizard Genomic DNA Purification Kit (Promega). Genome sequences of CF333-2007 and CF333-1991 were sequenced by 454 pyrosequencing with an approximate coverage of 20-fold. Reads were de novo assembled into contigs by Newbler version 1.1.03.24 (Roche) resulting in 164 and 87 contigs, respectively, that comprised 6,340,400 and 6,350,334 bp, respectively, and the two genome sequences was aligned using MUMmer3 (40) resulting in an alignment covering 6,336,193 bp of the genomes. Isolates CF114-1973, CF43-1973, CF66-1973, CF66-1992, CF66-2008, CF30-1979, CF173-1984, CF173-2005, CF333-1997, CF333-2003, and CF333-2007 were sequenced on an Illumina GAII generating 75-bp single-end reads using a multiplexed protocol to an average coverage of 47-fold (Table S2). The reads from each isolate were mapped against the reference genome of CF333-2007 using Bowtie release 0.12.3 (41), and SNP variants were called using SAMtools release 0.1.7 (42).

Construction of the Phylogenetic Tree. Genome sequences of all isolates show exhaustive homology to the reference CF333-2007 with no genome alignments covering less than 97.6% of the reference genome sequence (Table S2). Based on the SNP mutations identified in the alignments, we constructed a phylogenetic tree representing the evolutionary relationship among the DK2 strains. The branches at which the individual SNP mutations accumulated were determined by comparing the SNP mutations among the isolates. For example, if a mutation was found in isolates CF333-1997, CF333-2003, and CF333-2007, then the particular SNP mutation was assigned to the branch labeled "Q" in Fig. 2*A*. Except for two cases, all SNP mutations accumulated in a parsimonious fashion, such that a consistent tree could be constructed that shows the sequence of mutational events. This means that if an SNP mutation is shown to have accumulated on a particular branch, for example, branch K, then all clones descending from this branch (CF66-1992, CF66-2008, CF333-1991, CF333-1997, CF333-2003, and CF333-2007) contain the mutated allele. The only two SNP mutations that exhibit an aberrant evolutionary distribution that does not fit the observed sequential accumulation were found in *ampC* and *ftsI*. In the case of *ftsI*, a G188A substitution was found in isolates CF173-2005, CF66-1992, CF66-2008, CF333-1991, CF333-1997, CF333-2003, and CF333-2007 but not in CF173-1984 (Fig. 2*A*). The same ambiguity applies to *ampC*, in which an SNP mutation (A320G) was found in CF173-2005, CF66-1992, CF66-2008, CF333-2003, and CF333-2007 but not in CF333-1991, CF333-1997, and CF173-1984 (Fig. 2*A*). The irregular distribution of these particular SNP mutations may be a result of identical, but independent, mutational events, direct reversions, or horizontal transfer of DNA encoding the SNP. Notably, both *ampC* and *ftsI* have been reported to be under antibiotic-associated selection pressure (43, 44).

As the *ampC* and *ftsI* loci are the only two cases that confer ambiguity, we conclude that the tree structure is robust, although we cannot exclude the possibility that other similar events may be undetectable in the dataset. Because the immediate ancestor to the DK2 lineage is unavailable, a precise determination of the root of the tree is not possible. We predict the DK2 ancestor to be related to the sequenced strains as indicated by the predicted root in Fig. 2*A* based on the assumption that the SNP positions in the DK2 ancestor would be identical in sequence to the homolog sequences found in other *P. aeruginosa* reference strains such as PAO1, PA14, and LESB58.

Estimation of in Situ Bacterial Generations. The estimation of in situ generations of the DK2 lineage from 1973 to 2008 is based on our previous determination of in vivo growth rates of *P. aeruginosa* populations within sputum material sampled from different CF patients (15). Considering that growth rates of *P. aeruginosa* decrease when infection progresses in CF patients, we segregated the time period from 1973 to 2008 into several intervals and estimated the in

situ growth rate of each interval according to the in vitro growth rate. Generations of each isolates were calculated as sum of generations from 1973 to the isolation years (see details in Table S1 and *SI Materials and Methods*).

Global Phenotypic Profiling. Transcriptomes were measured using Affymetrix GeneChip for *P. aeruginosa*. Phenotype microarrays PM1, PM2, and PM3 (Biolog) were used to obtain the metabolic profiles of different carbon and nitrogen sources. Analysis of both types of datasets was performed using BioConductor software (see *SI Materials and Methods* for details) (45).

ACKNOWLEDGMENTS. We acknowledge the expert assistance of Francois Vigneault and George M. Church in genome sequencing. The project was supported by grants from the Danish Council for Independent Research/ Natural Sciences and from The Lundbeck Foundation.

1. Elena SF, Lenski RE (2003) Evolution experiments with microorganisms: The dynamics and genetic bases of adaptation. *Nat Rev Genet* 4:457–469.
2. Barrick JE, et al. (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* 461:1243–1247.
3. Shendure J, et al. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309:1728–1732.
4. Herring CD, et al. (2006) Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. *Nat Genet* 38: 1406–1412.
5. Cooper TF, Rozen DE, Lenski RE (2003) Parallel changes in gene expression after 20,000 generations of evolution in *Escherichia coli*. *Proc Natl Acad Sci USA* 100: 1072–1077.
6. Zambrano MM, Siegele DA, Almirón M, Tormo A, Kolter R (1993) Microbial competition: *Escherichia coli* mutants that take over stationary phase cultures. *Science* 259:1757–1760.
7. Giraud A, et al. (2008) Dissecting the genetic components of adaptation of *Escherichia coli* to the mouse gut. *PLoS Genet* 4:e2.
8. Rainey PB, Travisano M (1998) Adaptive radiation in a heterogeneous environment. *Nature* 394:69–72.
9. Boles BR, Thoendel M, Singh PK (2004) Self-generated diversity produces "insurance effects" in biofilm communities. *Proc Natl Acad Sci USA* 101:16630–16635.
10. Hansen SK, Rainey PB, Haagensen JA, Molin S (2007) Evolution of species interactions in a biofilm community. *Nature* 445:533–536.
11. Lyczak JB, Cannon CL, Pier GB (2002) Lung infections associated with cystic fibrosis. *Clin Microbiol Rev* 15:194–222.
12. Wilder CN, Allada G, Schuster M (2009) Instantaneous within-patient diversity of *Pseudomonas aeruginosa* quorum-sensing populations from cystic fibrosis lung infections. *Infect Immun* 77:5631–5639.
13. D'Argenio DA, et al. (2007) Growth phenotypes of *Pseudomonas aeruginosa* lasR mutants adapted to the airways of cystic fibrosis patients. *Mol Microbiol* 64:512–533.
14. Martin DW, et al. (1993) Mechanism of conversion to mucoidy in *Pseudomonas aeruginosa* infecting cystic fibrosis patients. *Proc Natl Acad Sci USA* 90:8377–8381.
15. Yang L, et al. (2008) In situ growth rates and biofilm development of *Pseudomonas aeruginosa* populations in chronic lung infections. *J Bacteriol* 190:2767–2776.
16. Jelsbak L, et al. (2007) Molecular epidemiology and dynamics of *Pseudomonas aeruginosa* populations in lungs of cystic fibrosis patients. *Infect Immun* 75: 2214–2224.
17. Winsor GL, et al. (2009) *Pseudomonas* Genome Database: Facilitating user-friendly, comprehensive comparisons of microbial genomes. *Nucleic Acids Res* 37(Database issue):D483–D488.
18. Nguyen D, Singh PK (2006) Evolving stealth: Genetic adaptation of *Pseudomonas aeruginosa* during cystic fibrosis infections. *Proc Natl Acad Sci USA* 103:8305–8306.
19. Rau MH, et al. (2010) Early adaptive developments of *Pseudomonas aeruginosa* after the transition from life in the environment to persistent colonization in the airways of human cystic fibrosis hosts. *Environ Microbiol* 12:1643–1658.
20. Huse HK, et al. (2010) Parallel evolution in *Pseudomonas aeruginosa* over 39,000 generations in vivo. *MBio* 1:e00199–10.
21. Naas T, Blot M, Fitch WM, Arber W (1994) Insertion sequence-related genetic variation in resting *Escherichia coli* K-12. *Genetics* 136:721–730.
22. Nghiem Y, Cabrera M, Cupples CG, Miller JH (1988) The mutY gene: A mutator locus in *Escherichia coli* that generates G.C—T.A transversions. *Proc Natl Acad Sci USA* 85: 2709–2713.
23. Smith EE, et al. (2006) Genetic adaptation by *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients. *Proc Natl Acad Sci USA* 103:8487–8492.
24. Kimura M (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ Press, Cambridge, UK).
25. Grocock RJ, Sharp PM (2002) Synonymous codon usage in *Pseudomonas aeruginosa* PA01. *Gene* 289:131–139.
26. Drake JW (1991) A constant rate of spontaneous mutation in DNA-based microbes. *Proc Natl Acad Sci USA* 88:7160–7164.
27. Ochman H, Elwyn S, Moran NA (1999) Calibrating bacterial evolution. *Proc Natl Acad Sci USA* 96:12638–12643.
28. Driffield K, Miller K, Bostock JM, O'Neill AJ, Chopra I (2008) Increased mutability of *Pseudomonas aeruginosa* in biofilms. *J Antimicrob Chemother* 61:1053–1056.
29. Ciofu O, Riis B, Pressler T, Poulsen HE, Høiby N (2005) Occurrence of hypermutable *Pseudomonas aeruginosa* in cystic fibrosis patients is associated with the oxidative stress caused by chronic lung inflammation. *Antimicrob Agents Chemother* 49: 2276–2282.
30. Boles BR, Singh PK (2008) Endogenous oxidative stress produces diversity and adaptability in biofilm communities. *Proc Natl Acad Sci USA* 105:12503–12508.
31. Oliver A, Canton R, Campo P, Baquero F, Blazquez J (2000) High frequency of hypermutable *Pseudomonas aeruginosa* in cystic fibrosis lung infection. *Science* 288: 1251–1254.
32. García-Castillo M, et al. (2010) Stationary biofilm growth normalizes mutation frequencies and mutant prevention concentrations in *Pseudomonas aeruginosa* from cystic fibrosis patients. *Clin Microbiol Infect*, 10.1111/j.1469-0691.2010.03317.x.
33. Barth AL, Pitt TL (1996) The high amino-acid content of sputum from cystic fibrosis patients promotes growth of auxotrophic *Pseudomonas aeruginosa*. *J Med Microbiol* 45:110–119.
34. Kureishi A, Diver JM, Beckthold B, Schollaardt T, Bryan LE (1994) Cloning and nucleotide sequence of *Pseudomonas aeruginosa* DNA gyrase gyrA gene from strain PAO1 and quinolone-resistant clinical isolates. *Antimicrob Agents Chemother* 38: 1944–1952.
35. Stover CK, et al. (2000) Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature* 406:959–964.
36. Dasgupta N, et al. (2003) A four-tiered transcriptional regulatory circuit controls flagellar biogenesis in *Pseudomonas aeruginosa*. *Mol Microbiol* 50:809–824.
37. Hentzer M, et al. (2003) Attenuation of *Pseudomonas aeruginosa* virulence by quorum sensing inhibitors. *EMBO J* 22:3803–3815.
38. Singh PK, et al. (2000) Quorum-sensing signals indicate that cystic fibrosis lungs are infected with bacterial biofilms. *Nature* 407:762–764.
39. Wiehlmann L, et al. (2007) Population structure of *Pseudomonas aeruginosa*. *Proc Natl Acad Sci USA* 104:8101–8106.
40. Kurtz S, et al. (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5:R12.
41. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.
42. Li H, et al. (2009) 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.
43. Spangenberg C, Montie TC, Tümmler B (1998) Structural and functional implications of sequence diversity of *Pseudomonas aeruginosa* genes oriC, ampC and fliC. *Electrophoresis* 19:545–550.
44. Gotoh N, Nunomura K, Nishino T (1990) Resistance of *Pseudomonas aeruginosa* to cefsulodin: Modification of penicillin-binding protein 3 and mapping of its chromosomal gene. *J Antimicrob Chemother* 25:513–523.
45. Gentleman RC, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5:R80.

Supporting Information

Yang et al. 10.1073/pnas.1018249108

SI Materials and Methods

Selection of Strains for Study. *Pseudomonas aeruginosa* isolates sampled from several hundred Danish cystic fibrosis (CF) patients since 1973 have been stored in the Copenhagen CF Clinic. The protocol for sampling and storage of isolates at the clinical laboratory was not designed with the purpose of sampling multiple isolates per patient per time point, and in most cases only single isolates were kept. Isolates were sampled and stored without prior knowledge of their particular genotype. Our molecular epidemiological analysis of the frozen collection of isolates showed that among all of the many different strains of *P. aeruginosa* isolated from the CF patients in Copenhagen, two were encountered repeatedly (1). Each of the two dominant strains, DK1 (formerly the "r" genotype) and DK2 (formerly the "b" genotype), have been isolated since 1973 from ∼40 patients. The major criteria for selecting strains for the present study were to ensure large temporal separation among the isolates, to ensure that isolates from several patients were included, and to ensure that multiple isolates from individual patients were selected. In the case of patient CF333, ∼20 strains isolated between 1991 and 2007 had been stored and were available for analysis. All isolates were determined to belong to the DK2 lineage, and for the present study we selected the first and last isolates (1991 and 2007) as well as two additional intermediate isolates (from 1997 and 2003). In the case of patient CF173, 26 isolates sampled between 1984 and 2005 were available, and we selected the first and last DK2 isolates. Similar to the clonal infection in patient CF333, DK2 strains were sampled continually from patient CF173 between 1984 and 2005, although from 2002 and onward both DK1 and DK2 isolates could be isolated simultaneously from the patient (1). The first isolate stored from patient CF66 was a DK2 strain sampled in 1973. Additional DK2 strains were not observed in this patient until the lineage reappeared in CF66 in 1992 because of transmission. All isolates sampled and stored between 1992 and 2008 were found to be DK2 strains. For the present study, we selected the 1973 isolate and the first (1992) and last (2008) isolates after reappearance. The CF30-1979 isolate was the only DK2 isolate from patient CF30. In the collection of 46 strains isolated from CF patients in 1973 (the earliest isolates available from the clinic), we identified four isolates of the DK2 genotype by means of genotyping and sequencing of single genes. One of the isolates (CF66-1973) had the same mutations in *mucA* and *rpoN* as all subsequent DK2 isolates from all analyzed patients. The three other DK2-1973 isolates had alternative *mucA* mutations and wild-type *rpoN* genes (two of the patients died in 1974 and 1975, respectively) and the DK2 clones went extinct, and the third patient died in 1991, after which the specific DK2 clone in this patient went extinct. There are no indications suggesting that any of these three DK2 clones were transmitted to other CF patients in the clinic. Thus, CF66-1973 appears to be the clinical ancestor of all subsequent DK2 isolates from more than 40 CF patients.

Genome Sequencing. Genomic DNA was prepared from *P. aeruginosa* isolates using a Wizard Genomic DNA Purification Kit (Promega). Genome sequences of CF333-2007 and CF333-1991 assembled from 454 pyrosequencing-generated data were aligned using MUMmer3 (2), resulting in the detection of 40 SNPs in the alignment covering 6,336,193 bp of the genomes. Genomes were sequenced with an approximate genomic coverage of 20-fold, and reads were assembled into contigs using Newbler version 1.1.03.24 (Roche). Genome sequences of CF333-1991 and CF333-2007 were assembled into 164 and 87 contigs, respectively, and comprised 6,340,400 and 6,350,334 base pairs, respectively. Twenty-eight of the SNPs were further confirmed by targeted PCR amplification of the region surrounding the SNP (∼300–500 bp) and subsequently resequenced at GATC Biotech. Isolates CF114-1973, CF43-1973, CF66-1973, CF66-1992, CF66-2008, CF30-1979, CF173-1984, CF173-2005, CF333-1997, CF333-2003, and CF333-2007 were sequenced on an Illumina GAII generating 75-bp single-end reads using a multiplexed protocol to an average coverage of 47-fold (Table S2). The reads from each isolate were mapped against the reference genome of CF333-2007 using Bowtie release 0.12.3 (3). Alignments were converted to Sequence Alignment/Map (SAM) format, and SNP variants were called using SAMtools release 0.1.7 (4). The consensus sequence was called according to the MAQ consensus model (5).

SNP Detection and Custom Filtering. The resulting raw SNP calls were filtered by the varFilter algorithm in SAMtools in which minimum SNP coverage was set to 3 followed by a custom filter. The filtered SNPs were further examined to have quality parameters equal to a subset of 28 SNPs that were confirmed by targeted PCR and resequencing.

Settings used in Bowtie and SAMtools for mapping and variant detection were as follows: bowtie -S <ref.fa> <reads.fastq> <out.sam>; samtools pileup -cv -f <ref.fa> <aln.bam>; samtools.pl varFilter -d 3 -D 10000. The resulting lists of detected SNPs were filtered in accordance with the following criteria: (*i*) The minimum Phred-scaled probability that the SNP call is identical to the reference was 30 (i.e., $P \leq 0.001$); (*ii*) ambiguous SNP (i.e., M, R, W, S, Y, and K) calls were removed; and (*iii*) SNPs must be called by no less than 50% of the covering reads. Subsequently, the called SNPs were subject to a manual filtering in which 70 SNPs were removed. These included 48 SNPs that were removed due to conflict of repetitive regions, 6 SNPs that emerged as a consequence of ungapped alignments near microindels, and 16 SNPs that displayed abnormal low-quality parameters in comparison with the subset of confirmed SNPs.

Estimation of the Fraction of SNPs in the Protein-Coding Part of the Genome That Results in Synonymous Substitutions. We employed the codon use in *P. aeruginosa* reference strain PAO1 (6) to calculate the probability that a random mutation would be synonymous assuming that the rates of different base changes are the same (7). We found the total number of possible SNP mutations within genes to be 16,779,042 (5,593,014 bp of coding sequences multiplied by 3 possible mutations in each position), of which 25% (4,237,247) would cause a synonymous change.

Estimation of in Situ Bacterial Generations. The estimation of in situ generations of the DK2 lineage from 1973 to 2008 is based on our previous determination of in vivo growth rates of *P. aeruginosa* populations in sputum material sampled from different CF patients (8). Using quantitative fluorescent in situ hybridization to measure cellular levels of rRNA, we have shown that in five sputum samples obtained in 2005, the majority (∼90%) of the *P. aeruginosa* lung population is actively growing with an average in vivo doubling time ranging from 115 to 141 min (8). One of the samples analyzed in our previous study was obtained from CF333 (previously designated "p7") in 2005 which harbored no *P. aeruginosa* strains other than DK2. The DK2 population in this sample had an average doubling time of 115 min in vivo,

which is about 1.5-fold longer than the measured in vitro doubling time during growth in LB.

We observed that earlier isolates have shorter in vitro generation times than later isolates (Table S1). The development of this slow-growth phenotype is most likely a consequence of the stressful CF environment in which the bacterial cells experience continuous exposure to antibiotics and contacts with the immune system. For example, growth rate-reducing effects have been observed for chromosomal mutations that confer antibiotic resistance (9–11). Because of these differences in in vitro doubling times, we therefore estimate the in vivo doubling times in different time periods as 1.5-fold the measured in vitro doubling times of the corresponding isolates during growth in LB. This extrapolation is based on the observation that growth rates of several isolates from 2005 including two DK2 strains consistently are 2- to 3-fold reduced compared with PAO1 or CF isolates from newly infected patients in LB as well as in various minimal media and sputum medium (8). To calculate the number of bacterial generations elapsed over the time period from 1973 to 2008, three intervals were defined: 1973 (CF66) to 1979 (CF30), 1979 (CF30) to 1997 (CF333), and 1997 to 2007 (CF333). The median doubling times of the isolates during in vitro growth in LB of each of the three intervals are 50, 56, and 74 min, respectively, from which the in vivo doubling times were estimated to be 74, 83, and 109 min, respectively. The number of generations for each isolate was calculated as the sum of generations from 1973 to the isolation year (Table S1). The total number of bacterial generations was estimated to calculate the mutation rate. The extrapolation from in vitro to in vivo doubling times is most likely a simplification. However, if we only apply the experimentally observed in vivo doubling time of 115 min measured in 2005 in a sample from CF333 and disregard the possibility that earlier isolates have faster in vivo growth rates, the total number of generations from 1973 to 2008 is estimated to be 159,965. The calculated mutation rate based on this number of bacterial generations is 9.4×10^{-11} SNPs/bp per generation, which is in the same range of the number calculated from 200,000 generations (7.2×10^{-11} SNPs/bp per generation).

Affymetrix GeneChip. For transcriptome analysis, all *P. aeruginosa* strains were grown aerobically at 37 °C in LB medium starting from OD_{600} = 0.01 and harvested at OD_{600} = 0.5. Harvested cells were mixed immediately with RNAprotect Bacteria Reagent (Qiagen) and stored at −80 °C. RNA was extracted with an RNeasy Mini Kit (Qiagen) and transcribed to cDNA using a SuperScript III Reverse Transcriptase (Invitrogen). The cDNA was purified with a Qiagen PCR Purification Kit (Qiagen) and fragmented using DNase I (Promega). Fragmented cDNA was then labeled using labeling reagent (Affymetrix) and terminal deoxynucleotidyl transferase (Promega). The labeled cDNA was hybridized on an Affymetrix *P. aeruginosa* PAO1 gene chip and stained on a GeneChip Fluidics Station 450. The probe arrays were finally scanned with a GeneChip Scanner 3000. The raw data (.cel files) were obtained using the Affymetrix GeneChip Operating System 1.4.

Biolog Phenotype Microarrays. *P. aeruginosa* strains were streaked on LB agar plates and incubated at 37 °C until colonies appeared on the plates (16–30 h). Cells were swabbed from the plates and suspended in IF-0 GN Base (inoculation fluid) at a density corresponding to 42% transmittance in the Biolog turbidimeter. The cell suspensions were diluted 1:6 in IF-0 minimal medium containing Biolog redox dye mixture D (tetrazolium), and 100-μL aliquots were added to carbon-source plates (PM1 and PM2A). For the nitrogen-source plate (PM3B), inoculations were supplemented with 30 mM glucose and 2 μM ferric citrate. The plates were incubated at 37 °C in an OmniLog plate reader (Biolog) for 72 h, and growth/respiration was measured kinetically by determining the colorimetric reduction of tetrazolium dye. Of the total 285 substrates available on the three PM plates, 159 were found informative (supporting consistent growth between replicates for at least one strain and not autoinducing the color change). Export of OmniLog data was performed using OmniLog *OL_FM/Kin* 1.20.02 software (Biolog). The average area beneath each kinetic curve was used for analysis. Total catabolic function was calculated as previously described (12). The growth capacity on 4-hydroxyphenylacetic acid (4-HPA) was confirmed in ABT minimal medium (8) supplemented with 10 mM 4-HPA as the sole carbon source. OD_{600} was measured after 60 h of growth at 37 °C.

Phenotypic Data Analysis. Analysis for both types of datasets was performed using BioConductor software (13). Normalization and probe set summary of the microarray expression data were performed with *RMA* (14). Hierarchical clustering was generated using the *hclust* function based on the Euclidean distance between strains using all genes measured on the Affymetrix GeneChips or all conditions measured on the Biolog phenotypic microarrays. The principal component analysis was conducted using singular-value decomposition with the *svd* function in R (15).

RpoN Complementation. Two isolates, CF333-1991 and CF333-2007, were complemented with a pMe6031 plasmid (16) that carried a wild-type copy of the *rpoN* gene through conjugation. Transformants were selected on pseudomonas isolation agar plates containing 60 μg/mL tetracycline.

Minimal Inhibitory Concentrations. Minimal inhibitory concentration values of ciprofloxacin were measured using Etest strips (AB Biodisk) by plating 100 μL of overnight culture on LB plates. The values were recorded after a 48-h incubation at 37 °C.

Sequencing of *lasR*, *mucA*, *algU*, and *rpoN* Genes. Sequencing of *lasR*, *mucA*, *algU*, and *rpoN* genes was performed by PCR using high-fidelity DNA polymerase (Fermentas) and four sets of primers: *lasR* forward (5′-TCTCGGACTGCCGTACAAC-3′) and *lasR* reverse (5′-AATTACCGATCGCCAGCTC-3′) for *lasR*, *mucA* fwd (5′-CTCTGCAGCCTTTGTTGCGAGAAG-3′) and *mucA* rev (5′-CTGCCAAGCAAAAGCAACAGGGAGG-3′) for *mucA*, *algU* fwd (5′-CCTGAGCCCGATGCAATCCATTTTCG-3′) and *algU* rev (5′-GGACAGAGTTTCCTGCAGGGCTTCAC-3′) for *algU* gene, and *rpoN* fwd (5′-CTCGCCAACGACCTGGTCAAG-3′) and *rpoN* rev (5′-GGCGTCGGTCACATCCAGTTG-3′) for *rpoN*. Another set of primers was used for *rpoN* sequencing: *rpoN* seq1 (5′-GCCTGGGAAGACATCTACCAGAC-3′) and *rpoN* seq2 (5′-GAGCTGAACCAGGAGGCGATG-3′).

Impact of *mucA*, *lasR*, and *rpoN* Mutations on Expression Profiles. To further evaluate the impact of the nonsynonymous mutations in *mucA*, *lasR*, and *rpoN* on the overall transcriptome changes of the late DK2 isolates, we compared the expression data of CF30-1979 carrying the three mutations with CF114-1973 without the three mutations (Table S1). In total, 1,005 genes were significantly different according to a Student's *t* test (two-tailed, $P < 0.01$). We conducted a similar analysis on a set of expression data of an intermittent CF isolate ($MucA^-$) and its isogenic $MucA^+$ pair from a previous study which was done under the same conditions as ours (17). There were 1,976 genes significantly changed due to the *mucA* mutation using the same statistical criteria. Between these two datasets, 471 genes overlapped with the same trend in terms of up- or down-regulation. In previous investigations, gene sets regulated by RpoN and LasR were also determined. Due to the lack of raw data of these two datasets, we could only use the filtered gene lists, which involved cutoffs of certain fold changes. Thus, these two datasets might have underestimated the gene numbers regulated by the two genes relative to the *mucA* mutation analysis. There were 62 genes regulated by RpoN (18) and 36 genes gov-

erned by LasR when cells were harvested at OD_{600} = 0.5 in LB medium (19). There were 19 and 22 genes, respectively, overlapping with the list of significant genes in the comparison of CF30-1979 and CF114-1973. Seventeen out of 19 genes controlled by RpoN were also in the list of the *mucA*$^-$ versus *mucA*$^+$ comparison. In total, among 1,005 significant genes comparing CF30-1979 against CF114-1973, 495 genes (49%) can be attributed to mutations in three genes, *mucA*, *rpoN*, and *lasR* (Fisher's exact test, $P = 1.77 \times 10^{-10}$), supporting the significance of the three regulatory genes in *P. aeruginosa* in CF airways.

Detection of the Genetic Marker Δ(PA1428–PA1431) in the DK2 Lineage. The *lasR* gene (PA1430) was removed in the DK2 lineage with a deletion of the region starting from the 342 bp of PA4128 and ending at 30 bp upstream of the *lasI* gene (PA1432). Using primers fwd (5′-AAAGGAATCCAGCATGCACATTCG-3′) and rev (5′-AGCCTTTCTGTCCAGAGTTGATGG-3′), a PCR product of 732 bp was generated for strains carrying the deletion, whereas an ∼5-kb fragment was produced for strains without the deletion. The deletion was used as a genetic marker to check the relation among 85 DK2 isolates from 42 patients (covering most of the patients who carried the DK2 strains in the Copenhagen CF Clinic, including those shown in Fig. 1). Since its first appearance in CF30-1979, this deletion was maintained in all analyzed isolates from 1984 to 2008 (no DK2 isolates were found between 1979 and 1984), which strongly suggests that all of the DK2 strains that have been successfully spreading in the clinic are descendants of CF30-1979.

Genetic Basis of Deficient Growth on 4-HPA in CF333-2007 and CF173-2005. Most of the genes involved in 4-HPA degradation in *P. aeruginosa* are located in an operon between PA4119 and PA4128 controlled by the transcriptional regulator encoded by PA4120 (20). The CF333-2007 isolate has a large gap [Δ(PA4116-87)] that deletes over 80 genes including PA4119–PA4128. Using primers fwd1 (5′-CCCAGGGCCTGAAGGAAGTG-3′) and rev1 (5′-CTGCTGCAGGTTCAGCGAAC-3′), a 912-bp fragment was amplified in CF333-2007 whereas no product was generated in CF30-1979. To exclude the latter amplification being a false negative, one gene in this region (PA4125) was amplified using primers PA4125 fwd (5′-GGCAGATCAACGCGGTATT-3′) and PA4125 rev (5′-GTTCAGGGTGCTGGCTGT-3′). CF173-2005 carries an ∼2-kb deletion (ΔPA4120-22) that removes PA4121 and partially PA4120 and PA4122. Using primers fwd2 (5′-GGTAGCCGGCAAGGTAGT-3′) and rev2 (5′-AGGGTCATGAAGCTGGAGAA-3′), a 579-bp DNA fragment was produced for CF173-2005 and an ∼2-kb fragment was amplified for CF30-1979.

***algU* Genotypes of 1973 Bacterial Isolates.** We note that the CF66-1973 isolate carries two mutations in the *algU* gene (Table S1). The A55G mutation, which is also present in later isolates, has previously (21) been shown not to neutralize the effects of the particular *mucA* ΔG430 mutation present in this and later isolates. In contrast, the other mutation, which is a 4-bp frameshift insertion, will counteract the effects of the *mucA* mutation. This particular *algU* mutation is not present in later isolates from this lineage, and most likely has been introduced during growth or the storage period in agar slants. Similar *algU* mutations are present in the other DK2 isolates from 1973, and because none of these carry the *rpoN* mutation found in the CF66 isolate, we have used the CF114 isolate as a reference with the highest resemblance to *P. aeruginosa* wild-type strain profiles.

1. Jelsbak L, et al. (2007) Molecular epidemiology and dynamics of *Pseudomonas aeruginosa* populations in lungs of cystic fibrosis patients. *Infect Immun* 75:2214–2224.
2. Kurtz S, et al. (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5:R12.
3. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.
4. Li H, et al.; 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.
5. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18:1851–1858.
6. Grocock RJ, Sharp PM (2002) Synonymous codon usage in *Pseudomonas aeruginosa* PA01. *Gene* 289:131–139.
7. Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* 15:496–503.
8. Yang L, et al. (2008) In situ growth rates and biofilm development of *Pseudomonas aeruginosa* populations in chronic lung infections. *J Bacteriol* 190:2767–2776.
9. Andersson DI, Levin BR (1999) The biological cost of antibiotic resistance. *Curr Opin Microbiol* 2:489–493.
10. Björkman J, Hughes D, Andersson DI (1998) Virulence of antibiotic-resistant *Salmonella typhimurium*. *Proc Natl Acad Sci USA* 95:3949–3953.
11. Björkman J, Nagaev I, Berg OG, Hughes D, Andersson DI (2000) Effects of environment on compensatory mutations to ameliorate costs of antibiotic resistance. *Science* 287:1479–1482.
12. Cooper VS, Lenski RE (2000) The population genetics of ecological specialization in evolving *Escherichia coli* populations. *Nature* 407:736–739.
13. Gentleman RC, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5:R80.
14. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19:185–193.
15. R Development Core Team (2007) R: A language and environment for statistical computing pages. *The R Project for Statistical Computing.* Available at http://www.R-project.org.
16. Heeb S, et al. (2000) Small, stable shuttle vectors based on the minimal pVS1 replicon for use in Gram-negative, plant-associated bacteria. *Mol Plant Microbe Interact* 13:232–237.
17. Rau MH, et al. (2010) Early adaptive developments of *Pseudomonas aeruginosa* after the transition from life in the environment to persistent colonization in the airways of human cystic fibrosis hosts. *Environ Microbiol* 12:1643–1658.
18. Dasgupta N, et al. (2003) A four-tiered transcriptional regulatory circuit controls flagellar biogenesis in *Pseudomonas aeruginosa*. *Mol Microbiol* 50:809–824.
19. Schuster M, Lostroh CP, Ogi T, Greenberg EP (2003) Identification, timing, and signal specificity of *Pseudomonas aeruginosa* quorum-controlled genes: A transcriptome analysis. *J Bacteriol* 185:2066–2079.
20. Zeng L, Jin S (2003) aph(3′)-IIb, a gene encoding an aminoglycoside-modifying enzyme, is under the positive control of surrogate regulator HpaA. *Antimicrob Agents Chemother* 47:3867–3876.
21. Ciofu O, et al.; Scandinavian Cystic Fibrosis Study Consortium (2008) Investigation of the algT operon sequence in mucoid and non-mucoid *Pseudomonas aeruginosa* isolates from 115 Scandinavian patients with cystic fibrosis and in 88 in vitro non-mucoid revertants. *Microbiology* 154:103–113.

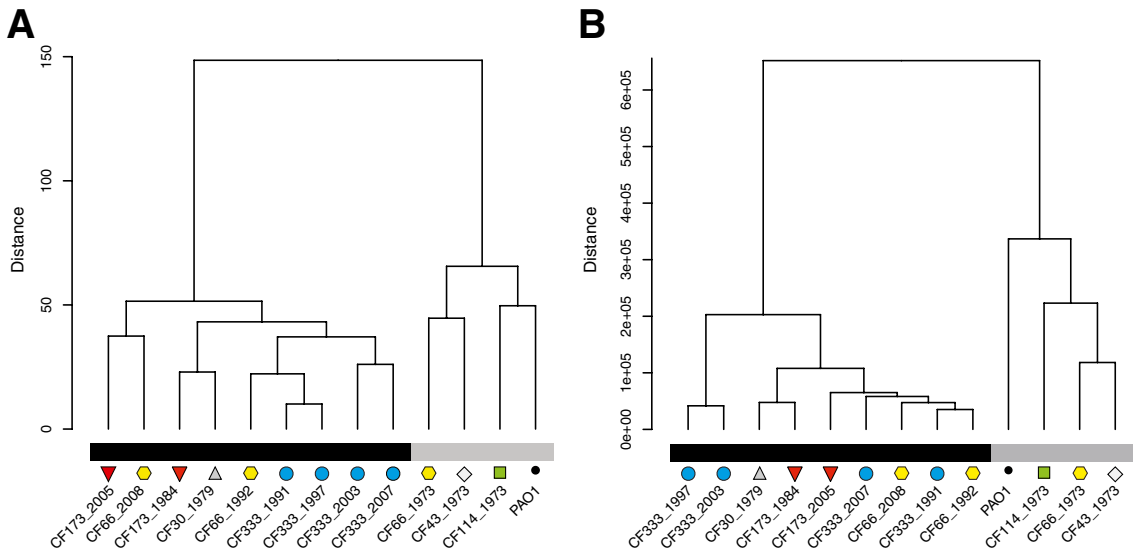


Fig. S1. Phenotypic relationships and dynamics among DK2 isolates. Dendrograms show the relationships among isolates based on microarray expression data (*A*) and Biolog phenotypic data (*B*). In both cases, the hierarchal clusters contain two major branches: one branch in which all isolates were sampled in/after 1979 (black bars), and another that groups the 1973 isolates together with reference strain PAO1 (gray bars). For each gene, average values of biological triplicates (*A*) or duplicates (*B*) were used for clustering. The hierarchical clustering dendrograms were constructed using the Euclidean distance.

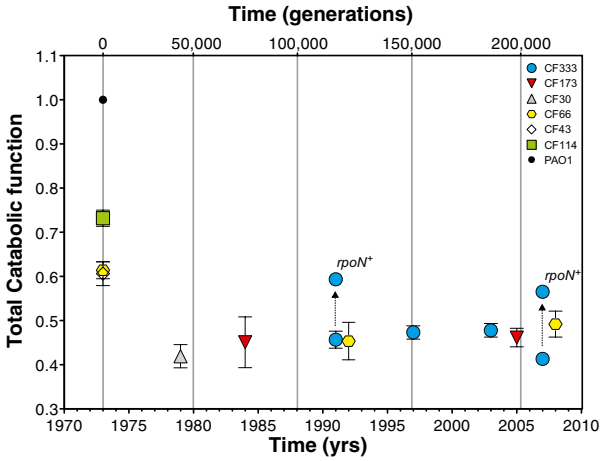


Fig. S2. Changes in catabolic functions of DK2 isolates relative to PAO1 as measured by Biolog phenotype microarrays. Total catabolic function is a weighted average across the 159 informative substrates for each strain; 1.0 is the reference level (PAO1), and lower values indicate decay. Values are shown on a log-transformed scale. DK2 isolates that were sampled after 1979 have a very low capacity of using various substrates compared with those from 1973; all changes were loss of function and no gains of function were observed. Complementation of CF333-1991 and CF333-2007 with a wild-type copy of *rpoN* largely restored the catabolic capacity.

Table S1. Selected phenotypes and gene sequences

Table S1 (DOC)

[a]CF strains are labeled with both a patient identifier and the year of isolation.
[b]In vitro doubling time was determined in LB liquid cultures. Numbers are average values of triplicates for each isolate (SD < 10%).
[c]3-O-C_{12}-homoserine lactone signal production controlled by *lasR* as well as motility phenotypes were measured as previously described (1).
[d]The effect of the *mucA* C382T mutations was neutralized by the *algU* C448G mutation because there was no induction of *algU* in the expression data of CF114-1973.
[e]The G647A mutation does not cause functional deficiency of LasR in CF66-1973 because the 3-O-C_{12}-homoserine lactone production is positive.

1. Lee B, et al. (2005) Heterogeneity of biofilms formed by nonmucoid *Pseudomonas aeruginosa* isolates from patients with cystic fibrosis. *J Clin Microbiol* 43:5247–5255.

Table S2. Illumina sequencing and mapping statistics

Table S2 (DOC)

[a]Reference genome CF333-2007 comprises 6,350,334 bp.

Table S3. SNPs in the DK2 lineage

Table S3 (DOC)

Table S4. Number of observed synonymous and nonsynonymous SNPs in protein-coding regions of the genome

Table S4 (DOC)

The expected number of nonsynonymous SNPs is calculated under the assumption of neutral selection, and the probability of seeing the observed number of nonsynonymous SNPs given the expected number of SNPs is calculated from the Poisson distribution. The ratio of the rate of nonsynonymous substitutions (dN) to the rate of synonymous substitutions (dS) was calculated as a measure of the selection pressure acting on the protein-coding genome. The ratio dN/dS is expected to exceed unity only if natural selection promotes changes in the protein sequence, whereas a ratio less than unity is expected only if natural selection suppresses protein changes.

Table S5. Nonsynonymous regulatory mutations that were fixed in the DK2 lineage

Table S5 (DOC)

Nonsynonymous regulatory mutations that were fixed in the DK2 lineage in CF333 until 2003 were calculated based on the list B, D, F, I, K, O, Q, and S of Fig. 2*A* and Table S3. In total, 16 nonsynonymous mutations occurred in regulators, and 11 of these arose before 1979. The annotation is based on *Pseudomonas* Genome database V2 (http://www.pseudomonas.com).
[a]Genes with deletions.

AAC Journals.ASM.org

Mutations in 23S rRNA Confer Resistance against Azithromycin in *Pseudomonas aeruginosa*

Rasmus Lykke Marvig,[a] Mette S. R. Søndergaard,[a] Søren Damkiær,[a] Niels Høiby,[b] Helle Krogh Johansen,[b] Søren Molin,[a] and Lars Jelsbak[a]

Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark,[a] and Department of Clinical Microbiology, Rigshospitalet and ISIM, University of Copenhagen, Copenhagen, Denmark[b]

The emergence of antibiotic-resistant *Pseudomonas aeruginosa* is an important concern in the treatment of long-term airway infections in cystic fibrosis patients. In this study, we report the occurrence of azithromycin resistance among clinical *P. aeruginosa* DK2 isolates. We demonstrate that resistance is associated with specific mutations (A2058G, A2059G, and C2611T in *Escherichia coli* numbering) in domain V of 23S rRNA and that introduction of A2058G and C2611T into strain PAO1 results in azithromycin resistance.

Systematic long-term, low-dose azithromycin treatment has been used for cystic fibrosis (CF) patients chronically infected with *Pseudomonas aeruginosa* in the Copenhagen CF center in Denmark since 2001 (3). Even though the use of azithromycin on CF lung infections has been found to have beneficial clinical effects (1, 3), it remains unclear how azithromycin works on *P. aeruginosa* and if macrolide resistance can emerge.

Using whole-genome sequencing, we have recently described the dissemination of the DK2 lineage of *P. aeruginosa* among a cohort of CF patients in Copenhagen, Denmark (15). The DK2 lineage has persisted for decades in the CF lung environment and has evolved as independent sublineages in concurrently infected patients. This makes it possible to search for convergent mutational events that occurred independently in isolates from different patients. Isolates CF333-2007 and CF66-2008 sampled from two different patients were found to have independently accumulated the same 23S rRNA A2045G mutation, while their most recent common ancestors did not (15). Mutations in the proximity of A2045 (position 2058 in *Escherichia coli* numbering) in the secondary structure of domain V of 23S rRNA have previously been shown to confer resistance toward macrolide antibiotics in other bacterial species (13). This led us to investigate if the observed mutations could be associated with azithromycin resistance and if such putative resistance mutations were found in strains of the DK2 lineage type present in other CF patients from the Copenhagen CF clinic.

We sequenced and analyzed the 23S rRNA genes of nine clinical *P. aeruginosa* isolates sampled from nine CF patients attending the Copenhagen CF center at the University Hospital (Table 1). Sequence data for three of the nine isolates were available from our previous study (15) (Table 1). Bacterial isolation from sputum and whole-genome sequencing were done as described previously (6, 15). Sequence reads were mapped against the 23S rRNA alleles of the *P. aeruginosa* DK2 reference genome (CF333-2007; GenBank accession no. CP003149), and single nucleotide polymorphisms in the 23S rRNA gene were called by SAMtools release 0.1.7 (10). The number of reads that supported a detected polymorphism relative to the total number of reads covering the polymorphic site was used to estimate the fraction of 23S rRNA genes in an individual clone which contained the polymorphism. Accordingly, polymorphisms present in one, two, three, or four of the four copies of 23S rRNA would be supported by approximately 25%, 50%, 75%, and 100% of the reads, respectively.

Three of the strains (CF206-2002, CF223-2002, and CF311-2002) had mutations in close proximity to the A2045 position in the secondary structure of domain V of 23S rRNA (13) at which isolates CF333-2007 and CF66-2008 were found to be mutated (Table 1). These mutations were A2046G (position 2059 in *E. coli* numbering) and C2598T (position 2611 in *E. coli* numbering), and both mutations have been reported to confer macrolide resistance in other bacterial species (13). The three remaining strains (CF180-2002, CF240-2002, and CF173-2005) did not contain any polymorphisms relative to the ancestral strain CF30-1979 in domain V of 23S rRNA.

We tested the resistance of the isolates toward azithromycin as described by Köhler et al. (9). Stationary-phase cultures of *P. aeruginosa* were incubated with and without 50 μg/ml azithromycin (16 to 62 μg/ml have been measured in CF sputum [14]) for 2 h, and after incubation, CFU were determined by plating onto LB agar plates with and without 20 μg/ml azithromycin, respectively. Inhibition of cell growth by azithromycin was calculated as the number of CFU from cultures grown with azithromycin relative to cultures grown without azithromycin.

As shown in Table 1, only minor or moderate effects of azithromycin treatment were observed for strains CF333-2007, CF66-2008, CF206-2002, CF223-2002, and CF311-2002 harboring mutations in domain V of 23S rRNA (15%, 4%, 25%, 12%, and 15% reduction in CFU, respectively). In contrast, strain CF30-1979, which closely resembles the ancestor of all DK2 clones isolated after 1979 (15) and harbors only wild-type 23S rRNA alleles, was found to be more inhibited by azithromycin (one-tailed Student's *t* test; $P < 0.05$) and exhibited a 64% reduction in CFU (Table 1).

Received 20 March 2012 **Returned for modification** 29 April 2012
Accepted 19 May 2012
Published ahead of print 29 May 2012
Address correspondence to Lars Jelsbak, lj@bio.dtu.dk.
R.L.M. and M.S.R.S. contributed equally to this article.

doi:10.1128/AAC.00630-12

TABLE 1 Resistance against azithromycin of clinical and laboratory strains used in this study

Strain	Mutations in 23S rRNA[b]	Fraction of reads supporting the mutations (%)	Start of systematic treatment with azithromycin (yr)	Resistance against azithromycin (%)[c]
Clinical isolates[a]				
CF30-1979[d]	None		2001	36 (±39)
CF333-2007[d]	A2045G (A2058G)	100	2006	85 (±14)
CF66-2008[d]	C1433T, A2045G (C1446T, A2058G)	21, 100	2001	96 (±39)
CF206-2002	C2598T (C2611T)	53	2006	75 (±32)
CF223-2002	C2598T (C2611T)	73	2001	88 (±35)
CF311-2002	A93G, A2046G (A106G, A2059G)	26, 70	2004	85 (±23)
CF180-2002	C1433T (C1446T)	25	2001	14 (±27)
CF240-2002	None		2006	35 (±32)
CF173-2005	None		2006	0
Laboratory strains				
PAO1	None	N/A[f]	N/A	0
PAO1-pMES-23S	None	N/A	N/A	1 (±1)
PAO1-pMES-23S(A2045G)	A2045G[e]	N/A	N/A	96 (±22)
PAO1-pMES-23S(C2598T)	C2598T[e]	N/A	N/A	92 (±14)

[a] Names of clinical isolates are given according to their patient origin and sampling year (i.e., isolate CF30-1979 was sampled from patient CF30 in 1979). All clinical isolates belong to the *P. aeruginosa* DK2 clone type.
[b] The corresponding *E. coli* 23S rRNA position is given in parentheses.
[c] Inhibition of cell growth by azithromycin was calculated as the number of CFU from cultures grown with azithromycin relative to cultures grown without azithromycin. The assay was replicated at least three times for each strain, and 95% confidence intervals are given in parentheses. No CFU appeared in cultures grown with azithromycin for strains PAO1 and CF173-2005.
[d] Whole-genome sequence of the isolate has been described previously (15).
[e] The strain carries a mutated 23S rRNA allele on the low-copy-number plasmid vector pME6031.
[f] N/A, not applicable.

The three strains (CF180-2002, CF240-2002, and CF173-2005) that did not contain any polymorphisms in domain V of 23S rRNA were all inhibited by azithromycin (one-tailed Student's *t* test; $P < 0.05$) (Table 1). Interestingly, the results also indicated that CF30-1979, CF180-2002, and CF240-2002 are less susceptible than PAO1. It is possible that this difference is due to other antibiotic-related mutations, e.g., in drug efflux pumps or ribosomal proteins L4 and L22 (8, 11, 12), present in the clinical strains.

To more directly assess the effects of the A2045G and C2598T mutations, we separately cloned both mutated and wild-type rRNA operons into the low-copy-number plasmid vector pME6031 (5 to 7 copies per cell) (4). The entire rRNA operon flanked by the genes PA4690 and PA4691 was PCR amplified from CF30-1979, CF333-2007, and CF223-2002 template DNA and ligated into plasmid pME6031, resulting in the complementation vectors pMES-23S, pMES-23S(A2045G), and pMES-23S(C2598T), respectively, which were introduced into the common laboratory strain *P. aeruginosa* PAO1 (Table 2).

Strain PAO1-pMES-23S was susceptible to azithromycin like its parent strain PAO1, whereas no inhibitory effect of azithromycin was observed for either PAO1-pMES-23S(A2045G) or PAO1-pMES-23S(C2598T) containing mutated 23S rRNA alleles, and both strains were significantly more resistant than PAO1-pMES-23S (one-tailed Student's *t* test; $P < 0.001$) (Table 1). This result confirmed the hypothesis that the A2045G and C2598T mutations confer resistance to azithromycin in *P. aeruginosa*.

We note that the resistance phenotype is observed both in strains PAO1-pMES-23S(A2045G) and PAO1-pMES-23S(C2598T) despite their capacity to express chromosomally encoded wild-type 23S rRNA and plasmid-encoded mutated 23S rRNA and in isolates CF206-2002, CF223-2002, and CF311-2002 in which none of the polymorphisms were fixed in all four copies of the 23S rRNA gene (Table 1). This indicates that not all four 23S rRNA alleles have to contain the mutations in order to confer the resistance phenotype (12).

With the exception of CF180, none of the other patients infected with susceptible isolates (CF30, CF240, and CF173) had been subjected to systematic long-term, low-dose treatment with azithromycin against *P. aeruginosa* at the time of sampling (Table 1). Moreover, resistant isolates CF333-2007, CF66-2008, and CF223-2002 were sampled from their respective host patients after initiation of systematic treatment with azithromycin. In con-

TABLE 2 List of plasmids and strains used for cloning purposes

Strain or plasmid	Relevant characteristics	Reference
E. coli CC118	Δ(*ara-leu*) *araD* Δ*lacX74 galE galK phoA20 thi-1 rpsE rpoB argE*(Am) *recAl*	5
P. aeruginosa PAO1		7
pME6031	PVS1-p15A shuttle vector; Tc^r	4
pMES-23S	A 5.95-kb SacI-PstI fragment containing the rRNA operon from CF30-1979 inserted into pME6031	This study
pMES-23S(A2045G)	A 5.95-kb SacI-PstI fragment containing the rRNA operon from CF333-2007 inserted into pME6031	This study
pMES-23S(C2598T)	A 5.95-kb SacI-PstI fragment containing the rRNA operon from CF223-2002 inserted into pME6031	This study

trast, the resistant isolates CF206-2002 and CF311-2002, both carrying mutations in domain V of 23S rRNA, were sampled from patients who had not been subjected to systematic treatment with azithromycin. It is possible that the presence of these resistant *P. aeruginosa* isolates is either the outcome of earlier treatments with macrolide antibiotics directed against infections caused by other CF pathogens (e.g., *Staphylococcus aureus*) in these patients or the result of transmission from patients treated with azithromycin.

In this study, we report the occurrence of azithromycin resistance among clinical *P. aeruginosa* isolates of the DK2 lineage. Sequencing of 23S rRNA genes revealed the resistance to be associated with mutations at specific loci (A2045, A2046, and C2598), and the effect of two of the mutations (A2045G and C2598T) were validated by *trans* complementation into the laboratory strain PAO1. Even though the use of azithromycin on *P. aeruginosa* CF lung infections has been found to have beneficial effects (1, 3), the exact mechanism by which azithromycin exerts this effect is not well understood. *In vitro*, the effect has been shown to be dependent on interaction with the ribosomes (2, 9), and our results suggest that the mode of action of azithromycin *in vivo* also involves binding to the 50S ribosomal subunit and that the action is blocked by mutations in 23S rRNA.

Further studies are required to investigate the diversity of resistance mutations, how different antibiotic treatment regimens influence the resistance development, and how the effect of azithromycin is dependent on the growth physiology of *P. aeruginosa*.

ACKNOWLEDGMENTS

We thank Morten Sommer for assistance in genome sequencing and helpful discussions and Katherine Long for helpful discussions and suggestions.

The project was supported by grants from the Danish Council for Independent Research/Natural Sciences and from The Lundbeck Foundation.

REFERENCES

1. **Cai Y, et al.** 2011. Effectiveness and safety of macrolides in cystic fibrosis patients: a meta-analysis and systematic review. J. Antimicrob. Chemother. **66**:968–978.
2. **Glansdorp FG, et al.** 2008. Using chemical probes to investigate the sub-inhibitory effects of azithromycin. Org. Biomol. Chem. **6**:4120–4124.
3. **Hansen CR, Pressler T, Koch C, Høiby N.** 2005. Long-term azithromycin treatment of cystic fibrosis patients with chronic Pseudomonas aeruginosa infection; an observational cohort study. J. Cyst. Fibros. **4**:35–40.
4. **Heeb S, et al.** 2000. Small, stable shuttle vectors based on the minimal pVS1 replicon for use in gram-negative, plant-associated bacteria. Mol. Plant Microbe Interact. **13**:232–237.
5. **Herrero M, de Lorenzo V, Timmis KN.** 1990. Transposon vectors containing non-antibiotic resistance selection markers for cloning and stable chromosomal insertion of foreign genes in gram-negative bacteria. J. Bacteriol. **172**:6557–6567.
6. **Høiby N, Frederiksen B.** 2000. Microbiology of cystic fibrosis, p 83–107. *In* Hodson M, Geddes D (ed), Cystic fibrosis, 2nd ed. Arnold, London, United Kingdom.
7. **Holloway BW, Krishnapillai V, Morgan AF.** 1979. Chromosomal genetics of Pseudomonas. Microbiol. Rev. **43**:73–102.
8. **Kataja J, Huovinen P, Skurnik M, The Finnish Study Group for Antimicrobial Resistance, Seppälä H.** 1999. Erythromycin resistance genes in group A streptococci in Finland. Antimicrob. Agents Chemother. **43**:48–52.
9. **Köhler T, Dumas JL, Van Delden C.** 2007. Ribosome protection prevents azithromycin-mediated quorum-sensing modulation and stationary-phase killing of Pseudomonas aeruginosa. Antimicrob. Agents Chemother. **51**:4243–4248.
10. **Li H, et al.** 2009. The sequence alignment/map format and SAMtools. Bioinformatics **25**:2078–2079.
11. **Ross JI, et al.** 1990. Inducible erythromycin resistance in staphylococci is encoded by a member of the ATP-binding transport super-gene family. Mol. Microbiol. **4**:1207–1214.
12. **Tait-Kamradt A, et al.** 2000. Mutations in 23S rRNA and ribosomal protein L4 account for resistance in pneumococcal strains selected in vitro by macrolide passage. Antimicrob. Agents Chemother. **44**:2118–2125.
13. **Vester B, Douthwaite S.** 2001. Macrolide resistance conferred by base substitutions in 23S rRNA. Antimicrob. Agents Chemother. **45**:1–12.
14. **Wilms EB, Touw DJ, Heijerman HG.** 2008. Pharmacokinetics and sputum penetration of azithromycin during once weekly dosing in cystic fibrosis patients. J. Cyst. Fibros. **7**:79–84.
15. **Yang L, et al.** 2011. Evolutionary dynamics of bacteria in a human host environment. Proc. Natl. Acad. Sci. U. S. A. **108**:7481–7486.

environmental microbiology

Environmental Microbiology (2012) **14**(8), 2200–2211 doi:10.1111/j.1462-2920.2012.02795.x

Deletion and acquisition of genomic content during early stage adaptation of *Pseudomonas aeruginosa* to a human host environment

Martin H. Rau,[1,2†] Rasmus Lykke Marvig,[1†] Garth D. Ehrlich,[3] Søren Molin[1,2*] and Lars Jelsbak[1]**

[1]*Department of Systems Biology, Technical University of Denmark, 2800 Lyngby, Denmark.*
[2]*Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Hørsholm, Denmark.*
[3]*Center for Genomic Sciences, Allegheny-Singer Research Institute, Pittsburgh, Pennsylvania, USA.*

Summary

Adaptation of bacterial pathogens to a permanently host-associated lifestyle by means of deletion or acquisition of genetic material is usually examined through comparison of present-day isolates to a distant theoretical ancestor. This limits the resolution of the adaptation process. We conducted a retrospective study of the dissemination of the *P. aeruginosa* DK2 clone type among patients suffering from cystic fibrosis, sequencing the genomes of 45 isolates collected from 16 individuals over 35 years. Analysis of the genomes provides a high-resolution examination of the dynamics and mechanisms of the change in genetic content during the early stage of host adaptation by this *P. aeruginosa* strain as it adapts to the cystic fibrosis (CF) lung of several patients. Considerable genome reduction is detected predominantly through the deletion of large genomic regions, and up to 8% of the genome is deleted in one isolate. Compared with *in vitro* estimates the resulting average deletion rates are 12- to 36-fold higher. Deletions occur through both illegitimate and homologous recombination, but they are not IS element mediated as previously reported for early stage host adaptation. Uptake of novel DNA sequences during infection is limited as only one prophage region was putatively inserted in one isolate, demonstrating that early host adaptation is characterized by the reduction of genomic repertoire rather than acquisition of novel functions. Finally, we also describe the complete genome of this highly adapted pathogenic strain of *P. aeruginosa* to strengthen the genetic basis, which serves to help our understanding of microbial evolution in a natural environment.

Received 15 April, 2012; revised 4 May, 2012; accepted 5 May, 2012. For correspondence. *E-mail sm@bio.dtu.dk; Tel. (+45) 45252513; Fax (+45) 45887328; **E-mail lj@bio.dtu.dk; Tel. (+45) 45256129; Fax (+45) 45932809. †These authors contributed equally to the work.

Introduction

Entry of a bacterium into a new environment provides an opportunity for expanding the possible niches it can inhabit. The inherent fitness of the bacterium towards the new environment as well as improvement of fitness by genetic adaptation will decide if entry is also followed by successful persistence. Modification of the genetic makeup of the organism may be the result of multiple mechanisms. For example, already present DNA can be changed by single nucleotide polymorphisms, small insertions and deletions and genome rearrangements. Also, larger genetic elements can be entirely lost or novel DNA may be gained by the processes of transformation, conjugation and transduction.

Depending on the environment certain mechanisms might be more prevalent than others. Acquisition of novel DNA can, e.g. be important for antibiotic resistance adaptation and acquisition of virulence enhancing elements for host infection (Hacker and Kaper, 2000; Campos-García, 2010). On the other hand adaptation towards permanent association with a host may be characterized by reductive evolution through extensive gene loss affecting functions that are no longer essential (Shigenobu *et al.*, 2000).

Studies of genome reduction are normally carried out either on present isolates compared with a theoretical ancestor or using *in vitro* evolutionary experiments. The former method is limited by a low resolution of the succession of events as the most recent common ancestor (MRCA) is usually millions of years older than the present day species that are investigated. *In vitro* evolution experiments allow for high resolution of the genomic change but the necessary time-scales are difficult to obtain and extrapolating directly from *in vitro* to *in vivo* should be done with caution.

Based on these two types of studies late stage reductive evolution is known to occur by pseudogene formation

and subsequent gradual loss of small genomic regions, whereas the process of reduction in the early stages is less clear with reports emphasizing either gene-by-gene non-functionalization (Silva *et al*., 2001; Dagan *et al*., 2006) or the loss of large contiguous regions (Moran and Mira, 2001; Nilsson *et al*., 2005).

For the lack of useful model organisms for *in vivo* monitoring of adaptation to a distinctly different environment, *Pseudomonas aeruginosa* infection of cystic fibrosis (CF) patients provides a useful alternative as well as an opportunity for studying the initial stages of host adaptation and the associated mechanisms of evolution. *Pseudomonas aeruginosa* is a highly versatile bacterium present in water and soil habitats but also capable of acute infection in a range of hosts from amoeba and plants to humans (Rahme *et al*., 2000; Pukatzki *et al*., 2002). Its vast array of functionalities also enables colonization of the lungs of CF patients, a colonization that is rapidly followed by genetic adaptation to be able to successfully persist in the new CF lung environment (Smith *et al*., 2006; Rau *et al*., 2010).

A specific clone of *P. aeruginosa*, DK2, with a history of infecting a large number of CF patients over at least 35 years was recently described by Yang and colleagues (2011a), providing an opportunity for studying the early stages of adaptation to a host environment. The initial study correlated the phenotypic change with a SNP-based phylogeny based on genome sequences of 12 *P. aeruginosa* DK2 isolates collected from six individuals over 35 years. Here, we describe the complete genome of this highly adapted pathogenic strain of *P. aeruginosa*. In addition, by sequencing the genomes of additional 33 DK2 clones sampled from 16 CF patients we have investigated the dynamics in the overall genomic content of the DK2 lineage, and focused on the role of genome reduction, acquisition of new DNA, and movement of mobile genetic elements in long-term adaptation to the new environment offered by its human hosts.

Results

Isolates of the DK2 clone type have been extensively collected at the Copenhagen CF centre at Rigshospitalet, Denmark, over a long time period ranging from 1973 and onwards. By sequencing the genomes of 45 DK2 isolates sampled from 16 individuals over a period of 35 years we have conducted a retrospective study of the dissemination of the *P. aeruginosa* DK2 clone type among CF patients attending the Copenhagen CF centre (Fig. 1). A previous study of DK2 strains by Yang and colleagues (2011a) dealt with the evolutionary relationship of isolates from six patients as determined by genome-wide SNP analysis based on a draft genome scaffold. To enable a more detailed investigation of the genetic content of this highly successful invader of the airways of CF patients, we completed and annotated the genomic features of clone CF333-2007, which has been sampled in 2007 from patient CF333 chronically infected with the DK2 strain for more than 16 years.

P. aeruginosa *DK2 genome features*

The entire genome sequence of *P. aeruginosa* DK2 clone CF333-2007a was completed using both Illumina and 454 reads in a hybrid *de novo* assembly strategy in which contigs were oriented based on the gene synteny in other *P. aeruginosa* strains. The CF333-2007a genome of *P. aeruginosa* DK2 (which we label PADK2) has a size of 6 402 658 bp comprising 5884 genes, which corresponds

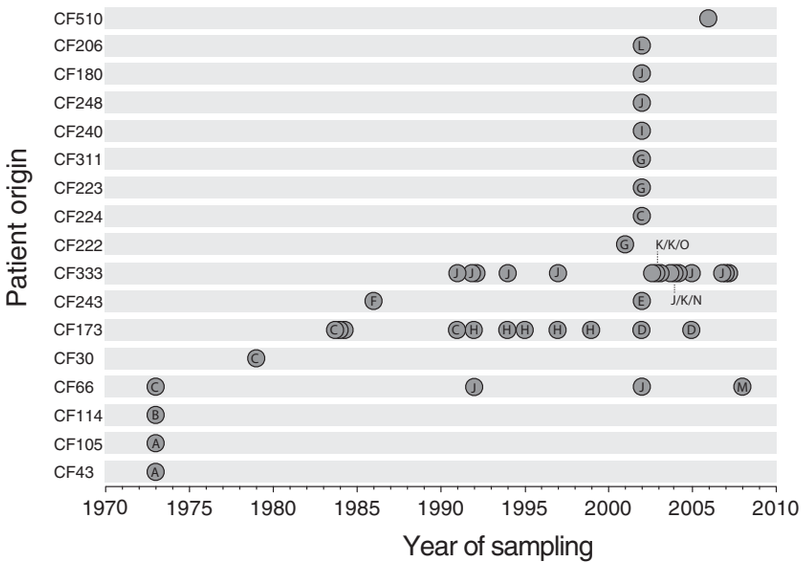


Fig. 1. Overview the patient origin and sampling time of the *P. aeruginosa* DK2 isolates used in this study. Bacterial isolates are represented by grey circles and their location in the phylogenetic tree in Fig. 3 is indicated by capital letters. If multiple isolates are sampled the same year from a patient, they are represented by stacked circles.

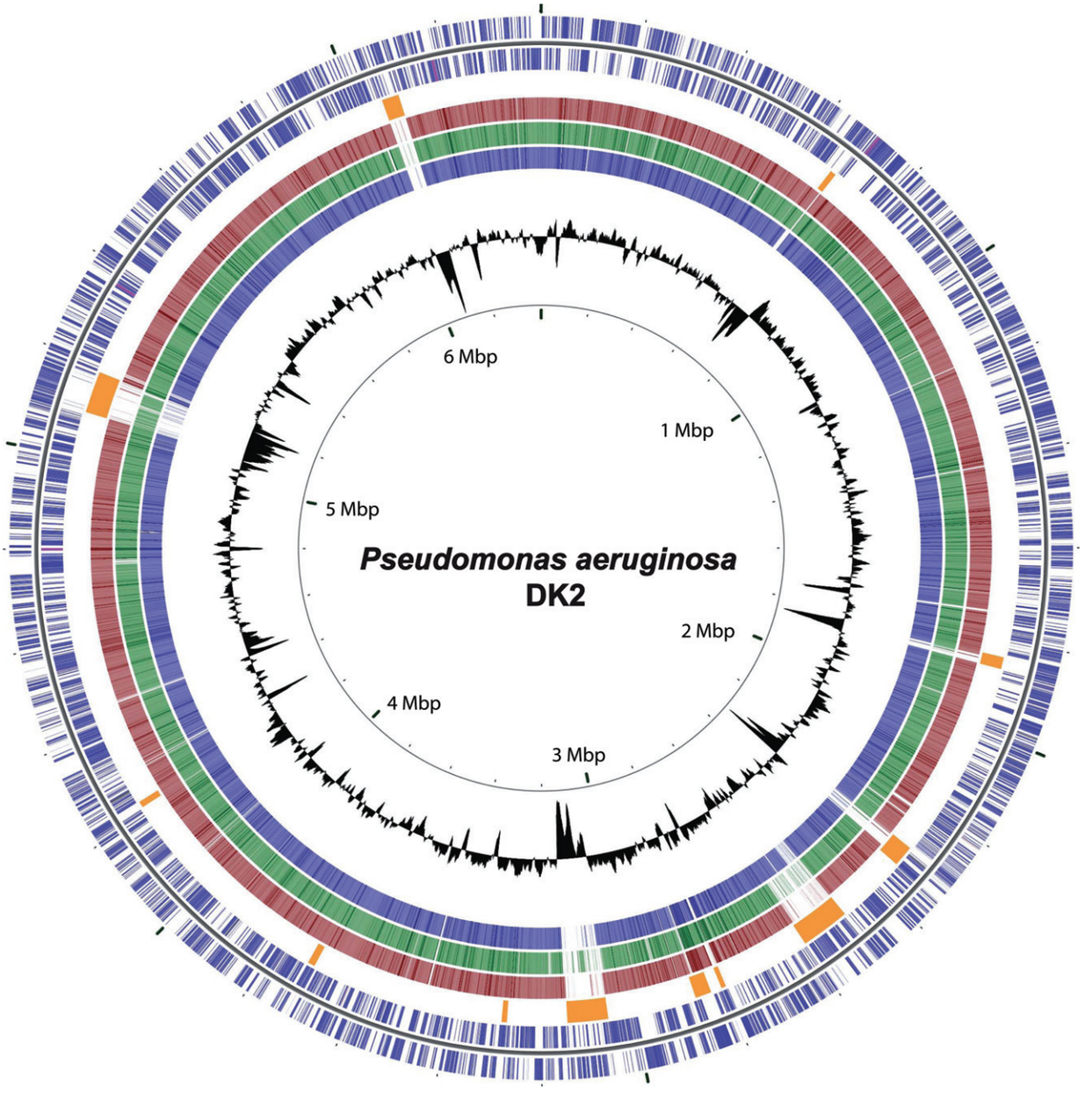


Fig. 2. Genome map of *P. aeruginosa* DK2. First and second outer circles denote CDS (blue) and rDNA (red); The third circle shows the location of genomic islands; the subsequent three rings display blast comparisons of the DK2 genome to *P. aeruginosa* PAO1 (red), PA14 (green), LESB (blue); the next ring depicts the GC content of the PADK2 genome.

to a coding density of 89.3% (Fig. 2). These numbers are similar to those of other assembled *P. aeruginosa* genomes as is the GC content of 66.3%. A large fraction of the genome is homologues to the other completed *P. aeruginosa* genomes of strains PAO1, LESB58, PA14 and PA7 (92.5%, 93.4%, 93.2% and 86.6% respectively) (Winsor *et al.*, 2011), whereas 216 kbp containing 195 genes were found to be unique to the DK2 strain relative to the other four strains.

The majority of this unique DNA (55%) was contained in three genomic islands related to heavy metal resistance (GI4, GI7 and GI12 respectively), indicating a former habitat of environmental origin with the presence of heavy metals. In total, there are 12 genomic islands (GIs) in the PADK2 genome constituting roughly 489 kbp (Table 1). The genome also contains two relatively small prophage regions, approximately 8 and 16 kbp each as well as eight different insertion sequence (IS) element types belonging

Table 1. Genomic islands present in the *P. aeruginosa* DK2 genome.

GI	Start	End	Size	Function[a]
GI1	667531	680249	12719	Prophage
GI2	1836969	1862655	25687	Serotype O3 O-antigen biosynthesis
GI3	2296150	2341644	45494	
GI4	2495033	2601031	105998	Copper resistance
GI5	2794504	2804831	10328	CRISPR region
GI6	2828700	2863808	35108	Pyoverdin biosynthesis
GI7	3056897	3144981	88084	Prophage, chromate/copper/silver resistance
GI8	3275228	3286591	11363	
GI9	3708420	3723852	15432	Type III secretion
GI10	4214902	4228275	13373	Non-ribosomoal peptide synthetase modules
GI11	5097794	5188335	90541	Type IVb pilus (PAGI-5-like GI)
GI12	6054401	6091531	37131	Arsenic resistance, nucleoside metabolism

a. The annotated function of a subset of genes within the GI.

to five different IS families. One IS element, IS222, is previously characterized while the other seven (ISPa36 to ISPa42) are new and added to the IS Finder Database (Siguier *et al*., 2006). Three IS elements belong to the TN3 family and all carry passenger genes (3, 6 and 14), while the other IS elements carry no passenger genes. Two CRISPR loci are present in the genome. One consists of two CRISPR regions flanking CRISPR-associated (cas) genes, while the other locus is devoid of cas genes. Apart from the spacers within the CRISPR regions, the former locus is identical to a CRISPR-cas system in *P. aeruginosa* strain PA14, where it has a different functionality than typical systems as it does not confer phage resistance but rather is mediating phage regulation of biofilm formation (Zegans *et al*., 2009; Cady *et al*., 2011).

Reductive evolution is prevalent in the DK2 lineage

Bacterial long-term adaptation to a host environment often includes reductive evolution in which non-essential functions are lost over time. We located all contiguous genome reductions of ≥ 1 kbp among 44 DK2 isolates sampled from chronically infected CF patients, and in a maximum-parsimonious phylogenetic manner we reconstructed the sequence of reductive events that was most likely to explain the observed genome reductions (Fig. 3).

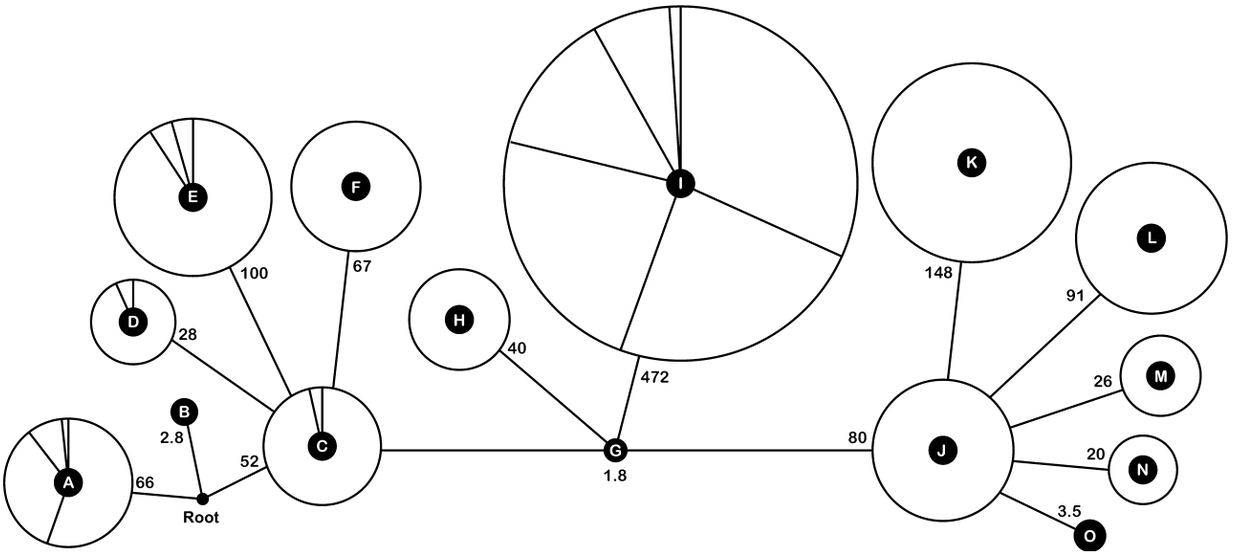


Fig. 3. Maximum-parsimonious phylogenetic tree reconstructing the sequence of large deletion events within the DK2 lineage. The tree is based on the genome sequences of 45 DK2 isolates which according to their genetic content locates in the nodes of the tree designated by capital letters. Refer to Table 3 or Fig. 1 to see the location of each isolate. The size of each node represents the total size (kbp) of genomic regions that have been deleted since the previous node. If more than one deletion has occurred, the circle is divided, and each subdivision corresponds to the size of each deletion (c.f. Table 2). The total quantity of deleted regions in an isolate represented by node J can for instance be found by summing all the deletion events along the path from the root to node J ($52 + 1.8 + 80 = 133.8$ kbp). The root represents the most recent common ancestor (MRCA) of all isolates.

Table 2. Description of deleted regions.

Node	Length	acc./core[a]	Function
A	1 160	acc.	
A	5 287	core	Exopolysaccharide (pslAB)
A	24 134	acc.	Chromate transport, arsenic resistance, copper resistance
A	35 534	core	
B	2 795	acc.	
C	1 455	core	Exopolysaccharide (pslFG)
C	50 834	acc.	Phage
D	1 957	core	Homoprotocatechuate metabolism (hpc)
D	26 610	acc.	RND efflux pump, silver efflux pump
E	4 332	core	
E	4 754	acc.	
E	91 043	acc.	PAGI-5
F	67 272	core	Homoprotocatechuate metabolism (hpc)
G	1 775	core	
H	40 295	core	Homoprotocatechuate metabolism (hpc)
I	7 156	core	Homoprotocatechuate metabolism (hpc)
I	38 588	acc.	
I	60 758	core	Virulence (pqsL, phzM/A1), biofilm (bfiRS), efflux (mexG-I)
I	91 043	acc.	PAGI-5
I	128 817	core/(acc.)	Iron uptake (pyoverdin)
I	145 995	core/(acc.)	Exopolysaccharide (pslA-I)
J	80 345	core	Homoprotocatechuate metabolism (hpc)
K	148 735	core/(acc.)	Adhesion (cupA1-5), cobalt/zinc cadmium resistance, virulence (katalase, ExoY toxin, hydrogen cyanide)
L	91 043	acc.	PAGI-5
M	26 104	core	Denitrification
N	20 141	acc./(core)	Iron uptake (pyoverdin)
O	3 532	core	

a. The column defines whether a deleted sequence belongs to core or accessory genome. If both types are noted both types are present within the sequence although the latter only in low percentage.

In total, 27 reductive events occurred during infection and since one specific deletion was observed thrice the 27 events comprised 25 different deletions (Table 2 and Table S1). The high consistency in the phylogeny (25/27 = 0.93) allows for an accurate depiction of the evolutionary relationship among the clones and predicted the clones to share a common ancestor located at the node between the branch point of the four earliest isolates sampled in 1973 (CF114-1973, CF105-1973, CF43-1973 and CF66-1973). The ancestor had a predicted approximate genome size of 6 535 548 bp. In total, 968 kbp (15%) of the genome were found to be deleted in at least one of the descendant DK2 clones sampled from chronically infected CF patients. Consequently, 968 kbp (~ 968 genes) of the ancestral *P. aeruginosa* DK2 genome is non-essential for survival in the human hosts.

The extent of reductive evolution varies significantly between isolates, but every isolate displays a reduced genomic content compared with the predicted ancestor. The level of genomic reduction is most extensive in CF240-2002 with a total reduction of 525 kbp since the MRCA corresponding to a loss of 8% of the entire genome. The majority of the reduction observed in CF240-2002 is specific to this isolate lineage only (472 kbp), the largest event being the loss of 146 kbp spanning mainly core genomic regions as well as minor accessory regions. The large deletions in CF240-2002 are caused by six separate events while most other reductions in individual isolates are the cause of only one event. The high number of large deletions in CF240-2002 could be stochastic, however, the genome of CF240-2002 contains missense mutations in two exonuclease genes, *sbcB* and *sbcC*, involved in recombination. In *E. coli*, mutation of *sbcB* enhances illegitimate recombination and increases the rate of deletion (Allgood and Silhavy, 1991), hence, the increased number of deletion events and the large quantity of lost regions in CF240-2002 could have a deterministic origin.

In total, 27 reductive events above 1 kbp are observed during the infection period with an average loss of 44.5 kbp per event (median 26.6 kbp). The lost genomic regions constitute both core regions and accessory regions, the latter containing, e.g. a pathogenicity island and genomic islands encoding heavy metal resistance genes, part of the DNA unique to the DK2 genome (Table 2). Overall, 56% of the lost regions belong to the core genome while 44% belong to the accessory genome, which is a significant over-representation considering the overall percentage of the accessory genome is 8.5%.

Deletion rates are high compared with in vitro *estimates*

Based on the knowledge of lost sequences and time points of strain isolation it is possible to obtain an estimation of the deletion rate during the reductive evolution. All DK2 clones sampled from chronic infections after 1973 share the same ancestor as recognized by proposed key-adaptive mutations in three global regulators (*mucA*, *lasR* and *rpoN*) and SNP-based phylogenetic analyses (Yang *et al.*, 2011a). The latter suggested clone CF30-1979, which is the earliest clone to possess all three key-adaptive mutations, to closely resemble this common ancestor of all isolates sampled from chronic infections from 1979 and onwards (Yang *et al.*, 2011a). In agreement with this, our analysis of genome reduction predicted CF30-1979 to represent an ancestor of all subsequently sampled clones (Fig. 3 and Table 3). Therefore, all deletions present in clones isolated after CF30-1979 most likely occurred after the year 1979, which therefore can be used as a reference point. The individual DK2 patient sub-lineages share evolutionary trajectory over several time periods before divergence, and the deletions of each patient sub-lineage are therefore not

Table 3. Location of isolates on the nodes in the phylogenetic tree in Fig. 3.

A:	CF114-1973	D:	CF173-2005	I:	CF240-2002	J:	CF333-2005
B:	CF43-1973	E:	CF243-2002	J:	CF66-1992	J:	CF333-2007a
B:	CF105-1973	F:	CF243-1986	J:	CF66-2002	J:	CF333-2007b
C:	CF66-1973	G:	CF222-2001	J:	CF180-2002	J:	CF333-2007c
C:	CF30-1979	G:	CF223-2002	J:	CF248-2002	K:	CF333-2003a
C:	CF173-1984a	G:	CF311-2002	J:	CF333-1991	K:	CF333-2003c
C:	CF173-1984b	H:	CF173-1992	J:	CF333-1992a	K:	CF333-2004b
C:	CF173-1984c	H:	CF173-1994	J:	CF333-1992b	L:	CF206-2002
C:	CF173-1991	H:	CF173-1995	J:	CF333-1994	M:	CF66-2008
C:	CF224-2002	H:	CF173-1997	J:	CF333-1997	N:	CF333-2004c
D:	CF173-2002	H:	CF173-1999	J:	CF333-2004a	O:	CF333-2003b

independent measures; however, the results will nonetheless provide a good indication of the deletion rate. Averaging the reduction rate per year for only the latest isolate for each patient (12 isolates) gives an overall deletion rate of 3723 bp per year. Estimating the deletion rate for the CF240-2002 sub-lineage gives a deletion rate of 22 637 bp per year.

Using previous estimates of *in vivo* doubling times (74–109 minutes), it is also possible to estimate the number of generations from 1979 to each end-point isolate (Yang *et al*., 2008; 2011a). A deletion rate of 3723 bp per year then corresponds to 0.62 bp per generation, which is 12- to 36-fold higher than previously reported *in vitro* deletion rates of 0.05 and 0.017 bp per generation for *Salmonella enterica* grown in rich medium (Nilsson *et al*., 2005). Although the extensive reduction observed in CF240-2002 significantly contributes to the high deletion rate, exclusion of this isolate from the calculations still result in a high rate of deletion (0.41 bp per generation).

For bacteria usually 20 bp of homologous sequence is required for homologous recombination (Ehrlich *et al*., 1993). Out of the 27 large deletions found here, seven of these contain homologous regions above 20 bp in length in the deletion end-point regions. The homologies of three of these are long and spans more than 100 bp. The remaining 20 deletions caused by illegitimate recombination have an average length of deletion end-point identity of 5.5 bp with 10 of the deletions having homologous regions below 5 bp.

Parallel reductive evolution

Two genomic regions have been lost in parallel for more than two sub-lineages during the infection period indicating a selective pressure against encoded functions in the region. In addition to this there are seven examples of overlap between two lost regions. One of the two parallel reductions is the fivefold parallel loss of genes encoding enzymes for homoprotocatechuate (hpc) catabolism, a sub-pathway in tyrosine catabolism. Although some of the five observed reductions also include the loss of large genomic regions surrounding hpc catabolism genes (up to 79 kbp) all five reductions include genes for this pathway making it the likely selective target of deletion.

The second example of parallel reductive evolution is the threefold loss of a region highly similar to the pathogenicity island PAGI-5, although the region in the DK2 isolates do not contain the main region of virulence genes (NR-II) identified in PAGI-5. This GI (GI11) spans 90 kbp and the threefold independent loss therefore constitutes a large fraction of the accessory genome reductions (50%). Two genomic islands closely related to PAGI-5 are pKLC102 and PAPI-1, both of which are mobilizable at respective frequencies of 10% and 0.16% (Qiu *et al*., 2006; Klockgether *et al*., 2007). Since PAGI-5 shares similar integration and transfer gene modules as PAPI-1 and pKLC102 it could possess similar mobilization frequencies. Accordingly, the threefold independent loss of the PAGI-5 homologue need not signify a selection against its presence. In fact, the preservation of the PAGI-5 homologue in many lineages could indicate that the GI contributes to fitness of the DK2 lineage.

We also note that deletion of different parts of the *psl* gene cluster occurred on three occasions (Table 2), which suggest selective pressure for loss of Psl exopolysaccharide production.

Acquisition of novel DNA

For rapid adaptation, acquisition of novel DNA can play a crucial role. However, during evolution of the DK2 lineage in the CF airways, acquisition of novel genetic elements, e.g. viruses or plasmids, seems to be less important in changing the genomic content. *De novo* assemblies of all genomes were performed to detect genetic elements not already present in the original infecting clone. Only 54 kbp of novel DNA sequence was found in CF114-1973 with half of it being homologous to phage D3 sequence indicating the insertion of a related prophage. However, since CF114-1973 is the first of the early 1973 clones to have diverged (Yang *et al*., 2011a), the sequence could have been present also in the ancestor and merely lost in the branch leading to subsequent isolates. No novel sequence could be found in other isolates.

The above is the only putative example of prophage insertion in the genomes of any DK2 strains during infection indicating a low level of temperate phage attack or high level of phage resistance in the DK2 population. Contrary, a GI containing two prophage regions was lost in the branch leading to CF66-1973 and all strains isolated subsequent to 1973. The spacers in the two CRISPR loci present in the DK2 lineage also remain unchanged for the duration of infection, indicating further the absence of temperate phage attack or simply that the systems are not functional. The identical system described by Cady and colleagues (2011) for strain PA14 was functional although 105 SNPs (within 8140 bp) differentiate it from the system in the DK2 strain.

Limited IS element expansion during infection

The *P. aeruginosa* DK2 genome contains eight different IS element types of which all exist in the 1973 isolates and have consequently not been acquired during infection. Only one type, IS222, was observed to differ in copy number in the DK2 isolates. IS222 existed in the five same loci in the genomes of CF114-1973, CF105-1973 and CF43-1973. During the branch leading to node C it was copied twice and is therefore present in seven loci in isolate CF66-1973 and all subsequent isolates. Furthermore, a duplication event of IS222 occurred in CF43-1973 hence having six copies, as well as in an ancestor to isolates CF66-1992, CF66-2002 and CF66-2008 leading to eight copies of IS222 in these isolates.

Discussion

The early stages of *P. aeruginosa* adaptation to the CF airways are characterized by the loss of large genomic regions. Whether the deletions are caused by drift or selection is unknown, but large deletions of core genome could facilitate the loss of beneficial genes, making the bacteria susceptible to Muller's ratchet. Although the largest proportion of reductions resides in the core genome there is a significant overrepresentation of reductive evolution in the accessory genome. This might also be expected as accessory functions often ensures proliferation in certain niches and here we, e.g. see the loss of some of the genes encoding heavy metal resistance in certain isolates. Probably, many of the accessory functions in the genome are neutral at best and therefore more prone to loss by drift or even selection against their functions.

The time scale in other studies of naturally evolved human pathogens compared with a known or theoretical ancestor is usually of a different magnitude spanning millions of years. *Burkholderia mallei* provides such an example with 1.4 Mbp lost over 3.5 million years (Song *et al*., 2010) and *Mycobacterium leprae* another with 408 out of 2977 genes deleted over 66 million years (Gomez-Valero *et al*., 2007). In these cases there is no information on the deletion dynamics. Consequently, due to a lack of intermediary strains rough estimates of *in vivo* deletion rates and dynamics have not been reported until now.

In vitro deletion rates have been reported for *Salmonella enterica* grown in rich medium rendering several biosynthesis functions superfluous (Nilsson *et al*., 2005). The observed deletion rates for the DK2 lineage is, however, 12- to 36-fold higher than those reported for the *S. enterica* evolutionary experiment. The *S. enterica* genome is around 4.8 Mbp while the *P. aeruginosa* DK2 genome is 6.4 Mbp, which could suggest that *P. aeruginosa* has a larger potential for genome reduction. In addition, *P. aeruginosa* is a versatile organism with a diverse metabolic repertoire, for which there is no need in the more confined human airways environment.

The CF lung contains a multitude of microbial species together with an active phage community, suggesting that this environment could constitute a rich source of new genomic material (Willner *et al*., 2009; Rolain *et al*., 2011). However, during the infection period of the DK2 lineage only little or no novel DNA was acquired. Only part of a prophage region was possibly acquired in CF114-1973. The prophage region was not present in other isolates, suggesting either that it entered in the branch leading to CF114-73 or was present in the shared ancestor but subsequently lost before the isolation of other 1973 strains. Even though the phage community in the CF lung has been shown to constitute a reservoir of, e.g. antibiotic resistance genes (Willner *et al*., 2009), no sign of impact of this is observed in the DK2 lineage. Possibly, the DK2 strains have reduced potential for uptake and acceptance of novel DNA as several receptors for DNA uptake, e.g. pili, flagella and LPS are either lost, reduced or altered during infection (Yang *et al*., 2011a,b). Investigating the phage community of explant lungs also revealed absence of phages in the highly diseased apical lobes, in contrast to other less diseased lung areas, suggesting phage resistance within that microbial community (Willner *et al*., 2012). Hence a possible feature of highly adapted *P. aeruginosa* strains in the CF lung is resistance towards phage predation. Regardless, since the DK2 clone emerges as a highly successful colonizer of the CF lung it stands to reason that lateral acquisition of DNA is in no way essential for *P. aeruginosa* adaptation to the CF lung and probably plays a limited role in this.

Transposable elements can impact evolution of an organism by inserting in new locations altering gene functionality or catalysing deletions by providing opportunity for homologous recombination. In the infecting lineage only one transposable element, IS222, seems capable of movement and the limited increase in copy number from

5 to 7 for most isolates, only affecting two genes, demonstrates its low significance in generating direct mutations compared with the high number of SNPs occurring in the genome over time (Yang *et al*., 2011a). This is in contrast to recent *in vitro* experiments revealing an important contribution of IS elements to mutation generation (Barrick *et al*., 2009; Gaffe *et al*., 2011). In the early stages of reductive evolution it has been suggested that extensive proliferation of IS elements occurs based on the high densities of these in fairly recent host-restricted species, such as *B. mallei*, *Bordetella pertussis*, *Shigella flexneri* and *Salmonella enterica* Typhi (Parkhill *et al*., 2003; Moran and Plague, 2004; Song *et al*., 2010). In *B. mallei* and *B. pertussis* there seems to be a link between the location of IS elements and large deletions and inversions, suggesting that IS elements function as catalysts for reductive evolution, and this mechanism was therefore proposed to be a key component in the early stage. However, at the very earliest point in adaptation of *P. aeruginosa* to the CF airways the proliferation of IS222 is still too limited for catalysis of reductions, which creates a potential for only very large (usually deleterious) reductions. Although the smallest distance between IS222 in later isolates is only 6 kbp the second-smallest is close to 340 kbp. Possibly, with a higher density of IS222 or the presence of another IS element with faster duplication the rate of reduction would be increased. Even though IS elements were estimated to be involved in nearly all large deletions in *B. mallei*, we see here in the earliest of stages substantial genome reduction of up to 8% of the genome without any involvement of IS elements.

Instead of IS mediated genome reduction we observe mechanisms of homologous recombination and more frequently, illegitimate recombination. Likely, this proportion is due to the abundance of non-homologous regions over homologous regions. Nilsson and colleagues (2005) also found mainly non-homologous end regions of reductions during *S. enterica in vitro* reductive evolution (6 regions) while one end region contained 40 bp with high identity when investigating the early stages of *in vitro* reductive evolution. These provide two examples of the earliest stage of host adaptation and both present reductive evolution through mainly illegitimate recombination.

The adaptation of the DK2 lineage to the CF lung environment provides for the first time insight into the early stages of adaptation of a bacterial pathogen to its human host. This early stage is exemplified by substantial reductive evolution predominantly through deletions of large genomic regions. Genomic change through uptake of novel DNA is on the other hand very limited, which is likely a reflection of gain of function being less important during transition to a permanent host association. The movement of mobile genetic elements is likewise scarce as IS element multiplication is low. Accordingly, IS elements are not involved in the process of genome reduction observed here. In spite of this high deletion rates are readily occurring through mechanisms of deletion mainly consisting of illegitimate recombination and to a lesser extent homologous recombination. Consequently, the involvement of IS elements is dispensable even for extensive genome reduction. Deletion rates are most likely determined by several factors, one of them being the genome reduction potential. This could explain the high deletion rate of the DK2 lineage as a significant portion of the coding capacity of *P. aeruginosa* should be geared towards its normal environmental habitats and therefore dispensable in the host environment.

Experimental procedures

Bacterial strains and genome sequencing

Forty-five isolates of the *P. aeruginosa* DK2 clone type were sampled over 35 years from 16 CF patients attending the Copenhagen Cystic Fibrosis Clinic at the University Hospital (Fig. 1). Isolation and identification of *P. aeruginosa* from sputum was done as previously described (Hoiby and Frederiksen, 2000). Sequencing of all isolates, except CF510-2006, was performed as earlier described (Yang *et al*., 2011a) on Illumina's GAIIx or Hiseq2000 platforms generating 75 base single-end reads using a multiplexed protocol to an average coverage depth of 11- to 99-fold (except CF333-1991 which was 454-sequenced only; cf. Yang *et al*., 2011a).

Assembly and annotation of the P. aeruginosa *DK2 reference genome*

The initial study of the *P. aeruginosa* DK2 clone type was based on a draft genome sequence of isolate CF333-2007a that was represented in 87 contigs assembled from 524 464 reads with an average length of 232 nucleotides generated from 454 pyrosequencing (Yang *et al*., 2011a). Isolate CF333-2007a was also sequenced on an Illumina GAII platform which generated 5 791 129 75-bp reads to complement the 454 generated reads and to increase the overall read coverage. By combining the two data sets into a hybrid assembly strategy, and by using knowledge of gene synteny in the genomes of other strains of *P. aeruginosa* genomes, we were able to complete the genome of CF333-2007a over several rounds of improvement:

First, the Illumina reads were *de novo* assembled with the assistance of genome sequences of *P. aeruginosa* PAO1, PA14, LESB58 and CF510-2006, respectively, using the Columbus module of Velvet version 1.0.16 (Zerbino and Birney, 2008). Each of reference-assisted *de novo* assemblies were then used to extend and close

gaps between the 87 original contigs using Minimus2 (Sommer *et al*., 2007), resulting in 49 new contigs.

We then manually inspected the 49 contigs and found all gaps to be caused by repetitive regions due to the following reasons: (i) copies of distant repeats collapsed onto one another into a single contig with ambiguous assemblies at the flanks, (ii) misassembly caused by collapse of tandem repeats, and (iii) incorrectly assembled chimeric contigs. For example, reads from the four nearly identical ribosomal RNA operons co-assembled into a single contig. Although this resulted in gaps in the *de novo* draft genome, we closed these gaps by first ordering the *de novo* contigs based on other available *P. aeruginosa* genome sequences followed by local re-assembly of the reads flanking the gaps. In the same way, gaps caused by tandem units of nearly identical repeats could be resolved by local re-assembly where only exact match reads were used, and chimeric contigs leaving orphan contig ends were broken up and re-assembled. A single gap consisting of 10 identical repeat units could not be resolved *in silico*, but was instead confirmed by PCR. In addition, the *in silico* gap closure was validated by PCR spanning 13 other gaps, including the gaps flanking the ribosomal RNA operons.

Finally, the Illumina reads was iteratively mapped back against the closed genome to correct small sequence errors, which would be identified as mutations. Pileups of the reads mapped by Novoalign (Novocraft Technologies) (Krawitz *et al*., 2010) were produced by SAMtools release 0.1.7 (Li *et al*., 2009), and single nucleotide polymorphisms (SNPs) called by the varFilter algorithm in SAMtools (samtools.pl varFilter -d 3 -D 10000 (–N 10) and SNP quality scores ≥ 50) were corrected in the genome sequence. In total, 531 nucleotides were corrected. The read alignments surrounding all putative indels were manually inspected, and 37 were corrected.

The final genome sequence was annotated using the NCBI Prokaryotic Genomes Automatic Annotation Pipeline (http://www.ncbi.nlm.nih.gov/genomes/static/Pipeline.html) and deposited at GenBank (Accession No. CP003149).

The accessory genome of *P. aeruginosa* DK2 was defined as genomic regions not present in at least two other fully sequenced *P. aeruginosa genomes* (strains PAO1, PA14, LESB, PA7) (Winsor *et al*., 2011). Alignment through MUMmer3 was used to locate accessory regions. Remaining regions were designated as core regions.

Genome sequencing and phylogenetic analysis of outgroup isolate CF510-2006

Clone CF510-2006 was isolated from a CF patient with an acute *P. aeruginosa* infection in 2006. Initial molecular genotyping using ArrayTube microarrays (Wiehlmann *et al*., 2007) showed that the isolate was genetically related to the DK2 clone type. However, the phenotype characteristics of the isolate were incompatible with contemporary DK2 strains present in chronically infected patients (Jelsbak *et al*., 2007; Yang *et al*., 2011a) which strongly suggest that the CF510-2006 isolate has little or no history of growth in the CF environment prior to its isolation. This is also supported by SNP profile comparison to DK2 isolates from Yang and colleagues (2011a), which show that CF510-2006 is more closely to clones isolated in 1973 than later isolates (see details below). The sequencing of the CF510-2006 genome was done by AGOWA (Berlin, Germany). First, 454 pyrosequencing reads were generated to an approximate genomic read-coverage of 50-fold that was assembled into 1205 contigs. Second, Sanger sequence reads from end-sequencing and primer walking of a fosmid genome library were used to close the majority of gaps to obtain a final genome containing 6 683 163 bp assembled into 71 contigs, which we annotated using the NCBI Prokaryotic Genomes Automatic Annotation Pipeline and deposited at GenBank under the accession AJHI00000000.

The genome of CF510-2006 was analysed with respect to the set of 363 SNPs that was previously identified in the reconstruction of the phylogenetic relationship of the *P. aeruginosa* DK2 lineage (Yang *et al*., 2011a). A total of 358 of these SNP-positions were present in the CF510-2006 genome, and were in all cases identical to the previously predicted MRCA of the DK2 lineage. The MRCA was predicted based on the assumption that its sequence at the informative SNP sites would be identical to the homologue sequences found in *P. aeruginosa* reference strains PAO1, PA14, LESB58 and PA7 (Winsor *et al*., 2011).

Analysis of genome reduction in multiple clones of the DK2 lineage

To examine the genetic content of each of the DK2 clones we first identified deletions larger than 1 kbp by alignment of reads against the genome sequence of CF510-2006. Hereby, regions not present in the clones could be detected (i.e. regions to which no reads aligned). Illumina reads from all isolates were mapped onto the genome of CF510-2006 using Novoalign (Novocraft Technologies) (Krawitz *et al*., 2010) and further processed by SAMtools release 0.1.7. A custom script was used to identify candidate regions larger than 1 kbp that was not present in the sequenced isolate as compared with CF510-2006. Subsequently, the candidate regions were manually inspected and exact deletion coordinates were verified by retrieving reads bridging the two regions adjacent to each deletion end. The genome sequence of isolate CF333-1991, which was assembled from 454 pyrosequencing reads, was

aligned to the CF510-2006 genome using MUMmer3 (Kurtz *et al*., 2004) and deleted regions were extracted from the alignment.

Some deletions spanned several contigs in CF510-2006 with unassembled sequence in between. For such cases, the corresponding contiguous sequence was identified in CF333-2007a and used instead based on the assumption that synteny would be preserved between the two isolates.

Mobile genetic elements

Genomic islands in the *P. aeruginosa* DK2 genome were detected by comparative genomics and prediction approaches. The first approach involved alignment to reference *P. aeruginosa* strains PAO1, PA14 and LESB58 genomes using MUMmer3 (Kurtz *et al*., 2004). Regions in the DK2 genome above 8 kbp and not present in two or more reference strains were designated as a genomic island. Additional genomic islands above 8 kbp were predicted using Islandviewer (Langille and Brinkman, 2009). Prophage regions in the DK2 genome were predicted using Prophinder (Lima-Mendez *et al*., 2008) using a normalized score of 0.5 as a cut-off, and Prophage Finder (Bose and Barber, 2006) using number of hits and predicted phage components as selection criteria. IS elements were located in the DK2 genome using ISsaga (Varani *et al*., 2011) (http://www-is.biotoul.fr). The expansion or loss of the IS elements during infection for Illumina sequenced isolates was tracked by analysis of reads bridging the IS element and its flanking regions. A change in location of an IS element or its multiplication was then detected as a change or increase in flanking regions while the loss of an IS element was detected as a loss of a flanking region.

The presence of CRISPRs in the DK2 genome was detected using CRISPRfinder (Grissa *et al*., 2007).

Deletion rates

Deletion rates per year were assessed for the most recently sampled isolate for each patient. The size in bp of deleted regions occurring since isolate CF30-1979 was divided by number of years from 1979 up until isolation year. The year 1979 and isolate CF30-1979 was chosen as reference time-point since CF30-1979 was isolated this year and based on SNP analysis (Yang *et al*., 2011a), it is very closely related to the ancestor of all isolates from 1979 and onwards.

Deletion rates per generation were calculated employing a previously described estimation of *in vivo* doubling times of the DK2 lineage (Yang *et al*., 2008; 2011a). These doubling times were used to estimate the number of generations from 1979 to the isolation year of the most recent isolate from each patient. Deletion rates per generation were then calculated using the number of generations from 1979 and the quantity of genome reduction that occurred after isolate CF30-1979.

Uptake of novel DNA

The genomes of Illumina sequenced isolates were *de novo* assembled using Velvet (version 1.0.17) (Zerbino and Birney, 2008) with the Columbus module providing reference assisted assembly. First, reads were mapped onto contigs of CF510-2006 using Novoalign (Novocraft Technologies) (Krawitz *et al*., 2010) and the assembly data were used during Velvet read assembly with CF510-2006 as reference. A coverage cut-off of 5 and a minimum contig length of 100 bp were used as parameters for velvetg while k-mer length of 35 was used for velveth. *De novo* assembled genomes were aligned to the genome of CF510-2006 using MUMmer3 (Kurtz *et al*., 2004) to locate novel sequence in each isolate. Any novel sequence not present in the genome of CF510-2006 was subsequently aligned to the genome of CF114-1973, to verify if the sequence was present in the earliest diverged isolate.

Acknowledgements

We are grateful to Nathan Ehrlich, Robert Boissy, Azad Ahmed, Ben Janto, and Justin Hogg from Center for Genomic Sciences, Allegheny-Singer Research Institute, for help with 454 sequencing and bioinformatics assistance related to the DK2 genome. The Lundbeck Foundation and the Danish Council for Independent Research (DFF) supported the work by grants to S.M.

References

Allgood, N.D., and Silhavy, T.J. (1991) Escherichia coli xonA (sbcB) mutants enhance illegitimate recombination. *Genetics* **127:** 671–680.

Barrick, J.E., Yu, D.S., Yoon, S.H., Jeong, H., Oh, T.K., Schneider, D., *et al*. (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* **461:** 1243–1247.

Bose, M., and Barber, R.D. (2006) Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. *In Silico Biol* **6:** 223–227.

Cady, K.C., White, A.S., Hammond, J.H., Abendroth, M.D., Karthikeyan, R.S., Lalitha, P., *et al*. (2011) Prevalence, conservation and functional analysis of *Yersinia* and *Escherichia* CRISPR regions in clinical *Pseudomonas aeruginosa* isolates. *Microbiology* **157:** 430–437.

Campos-García, J. (2010) Metabolism of acyclic terpenes by *Pseudomonas*. In *Pseudomonas*. Ramos, J.L., and Filloux, A. (eds). Dordrecht, the Netherlands: Springer Netherlands, pp. 235–253.

Dagan, T., Blekhman, R., and Graur, D. (2006) The 'domino theory' of gene death: gradual and mass gene extinction events in three lineages of obligate symbiotic bacterial pathogens. *Mol Biol Evol* **23:** 310–316.

Ehrlich, S.D., Bierne, H., d'Alencon, E., Vilette, D., Petranovic, M., Noirot, P., and Michel, B. (1993) Mechanisms of illegitimate recombination. *Gene* **135:** 161–166.

Gaffe, J., McKenzie, C., Maharjan, R.P., Coursange, E., Ferenci, T., and Schneider, D. (2011) Insertion sequence-driven evolution of *Escherichia coli* in chemostats. *J Mol Evol* **72:** 398–412.

Gomez-Valero, L., Rocha, E.P., Latorre, A., and Silva, F.J. (2007) Reconstructing the ancestor of *Mycobacterium leprae*: the dynamics of gene loss and genome reduction. *Genome Res* **17:** 1178–1185.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* **35:** W52–W57.

Hacker, J., and Kaper, J.B. (2000) Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol* **54:** 641–679.

Hoiby, N., and Frederiksen, B. (2000) Microbiology of cystic fibrosis. In *Cystic Fibrosis*. Hodson, M., and Geddes, D. (eds). London, UK: Arnold, pp. 83–107.

Jelsbak, L., Johansen, H.K., Frost, A.L., Thogersen, R., Thomsen, L.E., Ciofu, O., *et al*. (2007) Molecular epidemiology and dynamics of *Pseudomonas aeruginosa* populations in lungs of cystic fibrosis patients. *Infect Immun* **75:** 2214–2224.

Klockgether, J., Wurdemann, D., Reva, O., Wiehlmann, L., and Tummler, B. (2007) Diversity of the abundant pKLC102/PAGI-2 family of genomic islands in *Pseudomonas aeruginosa*. *J Bacteriol* **189:** 2443–2459.

Krawitz, P., Rodelsperger, C., Jager, M., Jostins, L., Bauer, S., and Robinson, P.N. (2010) Microindel detection in short-read sequence data. *Bioinformatics* **26:** 722–729.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol* **5:** R12.

Langille, M.G., and Brinkman, F.S. (2009) IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* **25:** 664–665.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., *et al*. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079.

Lima-Mendez, G., Van Helden, J., Toussaint, A., and Leplae, R. (2008) Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics* **24:** 863–865.

Moran, N.A., and Mira, A. (2001) The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. *Genome Biol* **2:** RESEARCH0054.

Moran, N.A., and Plague, G.R. (2004) Genomic changes following host restriction in bacteria. *Curr Opin Genet Dev* **14:** 627–633.

Nilsson, A.I., Koskiniemi, S., Eriksson, S., Kugelberg, E., Hinton, J.C., and Andersson, D.I. (2005) Bacterial genome size reduction by experimental evolution. *Proc Natl Acad Sci USA* **102:** 12112–12116.

Parkhill, J., Sebaihia, M., Preston, A., Murphy, L.D., Thomson, N., Harris, D.E., *et al*. (2003) Comparative analysis of the genome sequences of B*ordetella pertussis, Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat Genet* **35:** 32–40.

Pukatzki, S., Kessin, R.H., and Mekalanos, J.J. (2002) The human pathogen *Pseudomonas aeruginosa* utilizes conserved virulence pathways to infect the social amoeba *Dictyostelium discoideum*. *Proc Natl Acad Sci USA* **99:** 3159–3164.

Qiu, X., Gurkar, A.U., and Lory, S. (2006) Interstrain transfer of the large pathogenicity island (PAPI-1) of *Pseudomonas aeruginosa*. *Proc Natl Acad Sci USA* **103:** 19830–19835.

Rahme, L.G., Ausubel, F.M., Cao, H., Drenkard, E., Goumnerov, B.C., Lau, G.W., *et al*. (2000) Plants and animals share functionally common bacterial virulence factors. *Proc Natl Acad Sci USA* **97:** 8815–8821.

Rau, M.H., Hansen, S.K., Johansen, H.K., Thomsen, L.E., Workman, C.T., Nielsen, K.F., *et al*. (2010) Early adaptive developments of *Pseudomonas aeruginosa* after the transition from life in the environment to persistent colonization in the airways of human cystic fibrosis hosts. *Environ Microbiol* **12:** 1643–1658.

Rolain, J.M., Fancello, L., Desnues, C., and Raoult, D. (2011) Bacteriophages as vehicles of the resistome in cystic fibrosis. *J Antimicrob Chemother* **66:** 2444–2447.

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., and Ishikawa, H. (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* **407:** 81–86.

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* **34:** D32–D36.

Silva, F.J., Latorre, A., and Moya, A. (2001) Genome size reduction through multiple events of gene disintegration in *Buchnera* APS. *Trends Genet* **17:** 615–618.

Smith, E.E., Buckley, D.G., Wu, Z., Saenphimmachak, C., Hoffman, L.R., D'Argenio, D.A., *et al*. (2006) Genetic adaptation by *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients. *Proc Natl Acad Sci USA* **103:** 8487–8492.

Sommer, D.D., Delcher, A.L., Salzberg, S.L., and Pop, M. (2007) Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* **8:** 64.

Song, H., Hwang, J., Yi, H., Ulrich, R.L., Yu, Y., Nierman, W.C., and Kim, H.S. (2010) The early stage of bacterial genome-reductive evolution in the host. *PLoS Pathog* **6:** e1000922.

Varani, A.M., Siguier, P., Gourbeyre, E., Charneau, V., and Chandler, M. (2011) ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biol* **12:** R30.

Wiehlmann, L., Wagner, G., Cramer, N., Siebert, B., Gudowius, P., Morales, G., *et al*. (2007) Population structure of *Pseudomonas aeruginosa*. *Proc Natl Acad Sci USA* **104:** 8101–8106.

Willner, D., Furlan, M., Haynes, M., Schmieder, R., Angly, F.E., Silva, J., *et al*. (2009) Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE* **4:** e7370.

Willner, D., Haynes, M.R., Furlan, M., Hanson, N., Kirby, B., Lim, Y.W., *et al*. (2012) Case studies of the spatial heterogeneity of DNA viruses in the cystic fibrosis lung. *Am J Respir Cell Mol Biol* **46:** 127–131.

Winsor, G.L., Lam, D.K., Fleming, L., Lo, R., Whiteside, M.D., Yu, N.Y., *et al*. (2011) Pseudomonas Genome Database: improved comparative analysis and population genomics capability for Pseudomonas genomes. *Nucleic Acids Res* **39:** D596–D600.

Yang, L., Haagensen, J.A., Jelsbak, L., Johansen, H.K., Sternberg, C., Hoiby, N., and Molin, S. (2008) *In situ* growth rates and biofilm development of *Pseudomonas aeruginosa* populations in chronic lung infections. *J Bacteriol* **190:** 2767–2776.

Yang, L., Jelsbak, L., Marvig, R.L., Damkiaer, S., Workman, C.T., Rau, M.H., *et al*. (2011a) Evolutionary dynamics of bacteria in a human host environment. *Proc Natl Acad Sci USA* **108:** 7481–7486.

Yang, L., Rau, M.H., Hoiby, N., Molin, S., and Jelsbak, L. (2011b) Bacterial adaptation during chronic infection revealed by independent component analysis of transcriptomic data. *BMC Microbiol* **11:** 184.

Zegans, M.E., Wagner, J.C., Cady, K.C., Murphy, D.M., Hammond, J.H., and O'Toole, G.A. (2009) Interaction between bacteriophage DMS3 and host CRISPR region inhibits group behaviors of *Pseudomonas aeruginosa*. *J Bacteriol* **191:** 210–219.

Zerbino, D.R., and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18:** 821–829.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Table S1. The table lists the genomic coordinates of the deletions and whether they are detected in an isolate (YES denotes that a deletion is detected).

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

OPEN ACCESS Freely available online

PLOS | GENETICS

Genome Analysis of a Transmissible Lineage of *Pseudomonas aeruginosa* Reveals Pathoadaptive Mutations and Distinct Evolutionary Paths of Hypermutators

Rasmus Lykke Marvig[1], Helle Krogh Johansen[1,2], Søren Molin[1]*, Lars Jelsbak[1]*

1 Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark, **2** Department of Clinical Microbiology, Rigshospitalet, Copenhagen, Denmark

Abstract

Genome sequencing of bacterial pathogens has advanced our understanding of their evolution, epidemiology, and response to antibiotic therapy. However, we still have only a limited knowledge of the molecular changes in *in vivo* evolving bacterial populations in relation to long-term, chronic infections. For example, it remains unclear what genes are mutated to facilitate the establishment of long-term existence in the human host environment, and in which way acquisition of a hypermutator phenotype with enhanced rates of spontaneous mutations influences the evolutionary trajectory of the pathogen. Here we perform a retrospective study of the DK2 clone type of *P. aeruginosa* isolated from Danish patients suffering from cystic fibrosis (CF), and analyze the genomes of 55 bacterial isolates collected from 21 infected individuals over 38 years. Our phylogenetic analysis of 8,530 mutations in the DK2 genomes shows that the ancestral DK2 clone type spread among CF patients through several independent transmission events. Subsequent to transmission, sub-lineages evolved independently for years in separate hosts, creating a unique possibility to study parallel evolution and identification of genes targeted by mutations to optimize pathogen fitness (pathoadaptive mutations). These genes were related to antibiotic resistance, the cell envelope, or regulatory functions, and we find that the prevalence of pathoadaptive mutations correlates with evolutionary success of co-evolving sub-lineages. The long-term co-existence of both normal and hypermutator populations enabled comparative investigations of the mutation dynamics in homopolymeric sequences in which hypermutators are particularly prone to mutations. We find a positive exponential correlation between the length of the homopolymer and its likelihood to acquire mutations and identify two homopolymer-containing genes preferentially mutated in hypermutators. This homopolymer facilitated differential mutagenesis provides a novel genome-wide perspective on the different evolutionary trajectories of hypermutators, which may help explain their emergence in CF infections.

Citation: Marvig RL, Johansen HK, Molin S, Jelsbak L (2013) Genome Analysis of a Transmissible Lineage of *Pseudomonas aeruginosa* Reveals Pathoadaptive Mutations and Distinct Evolutionary Paths of Hypermutators. PLoS Genet 9(9): e1003741. doi:10.1371/journal.pgen.1003741

Editor: Diarmaid Hughes, Uppsala University, Sweden

Received May 13, 2013; **Accepted** July 8, 2013; **Published** September 5, 2013

Funding: This work was supported by the Danish Council for Independent Research (http://fivu.dk/forskning-og-innovation/rad-og-udvalg/det-frie-forskningsrad) and the Lundbeck Foundation (http://www.lundbeckfonden.com). HKJ was supported by a clinical research stipend from the Novo Nordisk Foundation (http://www.novonordiskfonden.dk). LJ was supported by the Villum Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: sm@bio.dtu.dk (SM); lj@bio.dtu.dk (LJ)

Introduction

A molecular and mechanistic understanding of how bacterial pathogens evolve during infection of their human hosts is important for our ability to fight infections. The advent of high-throughput sequencing techniques now offer unprecedented nucleotide resolution to determine the relatedness among infecting bacterial isolates and to unveil genetic adaptation within infected individuals and in response to antibiotic therapy [1–7]. Unraveling the genetic content of pathogens helps to identify the genes that make certain bacterial lineages more pathogenic than others. Nonetheless, the pathogenicity of a bacterial clone can also evolve via the mutational changes of pre-existing genes, a mechanism which is also known as pathogenicity- or pathoadaptive mutations [8]. While several studies have provided insight into the genomic evolution of primary bacterial pathogens such as *Yersinia pestis* [6] and *Vibrio cholera* [2] causing acute infections, only little is known how these observations relate to opportunistic pathogens causing long-term infections [1].

The opportunistic pathogen *Pseudomonas aeruginosa* is a common environmental inhabitant, which is also capable of causing both acute and chronic infections in a range of hosts from amoeba and plants to humans. For example, *P. aeruginosa* causes chronic airway infections in most patients with cystic fibrosis (CF), and is directly associated with the morbidity and mortality connected with this disease. Chronic CF infections provide an opportunity for long-term monitoring of the battle between the infecting bacteria and the host immune defense and clinical intervention therapy [9,10], and thus offer a direct method for observing evolutionary mechanisms *in vivo*. In an effort to understand the evolutionary mechanisms facilitating the transition of *P. aeruginosa* from its environment to a human host, we have previously found no

Author Summary

Pseudomonas aeruginosa is the dominating pathogen of chronic airway infections in patients with cystic fibrosis (CF). Although bacterial long-term persistence in CF hosts involves mutation and selection of genetic variants with increased fitness in the CF lung environment, our understanding of the within-host evolutionary processes is limited. Here, we performed a retrospective study of the *P. aeruginosa* DK2 clone type, which is a transmissible clone isolated from chronically infected Danish CF patients over a period of 38 years. Whole-genome analysis of DK2 isolates enabled a fine-grained reconstruction of the recent evolutionary history of the DK2 lineage and an identification of bacterial genes targeted by mutations to optimize pathogen fitness. The identification of such pathoadaptive genes gives new insight into how the pathogen evolves under the selective pressures of the host immune system and drug therapies. Furthermore, isolates with increased rates of mutation (hypermutator phenotype) emerged in the DK lineage. While this phenotype may accelerate evolution, we also show that hypermutators display differential mutagenesis of certain genes which enable them to follow alternative evolutionary pathways. Overall, our study identifies genes important for bacterial persistence and provides insight into the different mutational mechanisms that govern the adaptive genetic changes.

evidence for horizontal acquisition of genes to play a role [9]. Instead, we suggested the establishment of long-term chronic infections to be a matter of tuning the existing genome via pathoadaptive mutations.

The within-host mutation rate is a key factor in determining the potential for bacterial pathogens to genetically adapt to the host immune system and drug therapies, and knowledge about *in vivo* growth dynamics of bacterial pathogens and their capacity for accumulation of mutations is essential for the design of optimal interventions. Interestingly, the generation of mutations is frequently accelerated in clinical populations of *P. aeruginosa* that evolves as so-called 'hypermutators' due to deficient DNA mismatch repair systems [11]. Although the hypermutable phenotype is also observed for other species in a range of conditions [12–16], the impact of this phenotype in a natural environment and in relation to infections remains less clear.

Here we analyze the genome sequences of 55 isolates of the transmissible *P. aeruginosa* DK2 clone type causing chronic infections in a cohort of Danish CF patients. Our collection, comprising both normal (normomutator) and hypermutable isolates, enabled a comparative analysis of evolutionary trajectories of individual sub-lineages of the DK2 clone type making it possible to identify genes targeted by pathoadaptive mutations. Furthermore, the long-term population dynamics and structure of the clonal expanding DK2 lineage was elucidated by a high-resolution phylogeny, and an examination of the mutation dynamics of homopolymers (homopolymeric tracts of identical nucleotides, *e.g.* GGGGG) provided novel genome-wide evidence for the potential advantage of differential mutagenesis associated with the hypermutator phenotype.

Results and Discussion

Strain collection and maximum-parsimonious phylogeny

We sequenced the genomes of a collection of 55 *P. aeruginosa* DK2 clones sampled from Danish CF patients between 1972 and 2008 (Figure 1 and Table S4). The sequence data of 45 of the isolates have previously been reported [9,10]. Most patients (n = 19) were represented by only a single or a few (≤4) isolates. However, two patients were represented by 11 and 15 isolates (CF173 and CF333, respectively).

We identified a total of 7,326 unique SNPs in the 55 DK2 genomes, that could be explained by 7,368 mutational events (consistency index 0.99) using a maximum-parsimonious phylogenetic model to elucidate the evolutionary relationship of the *P. aeruginosa* DK2 population (Figure 2). The high consistency of the tree reflects the unidirectional, clonal evolution from the root of the tree to the tips, thus enabling inferences about the succession of mutations and the relationship among *P. aeruginosa* DK2 clones.

Prevalence of mutS, mutL, and mutY hypermutators

From the phylogenetic tree we observed a linear correlation between the number of SNPs and the time of sampling (*i.e.* a constant rate of mutation accumulation during the clonal expansion of the DK2 lineage) (Figure 3). However, nine sub-lineages (indicated by filled circles in Figure 3) deviated from this trend and had accumulated mutations at higher rates. In one of these isolates, CF224-2002a, we found that 265 of the 273 SNPs accumulated in the branch leading to the isolate were densely clustered in two chromosomal regions with SNP densities (1.2 and 1.8 SNPs per kb, respectively), that are much higher than expected (0.043 SNPs per kb assuming a random distribution of SNPs) (Figure S1). The most likely explanation for these high SNP densities is that the two genomic regions are the result of recombination events with DNA from a *P. aeruginosa* strain(s) unrelated to the DK2 clone type. Another study by Chung *et al.* observed similar indications of within-patient recombination events in *P. aeruginosa* [17].We found no evidence for additional recombination events among the 55 genome sequences.

The excess numbers of mutations in the remaining eight deviant isolates were the result of increased mutation rates due to mutations in mismatch repair and error prevention genes. Seven isolates had non-synonymous mutations in one of the DNA mismatch repair (MMR) genes *mutS* (n = 2) and *mutL* (n = 4) or both (n = 1), and their excess numbers of SNPs showed a highly increased transition:transversion ratio consistent with MutS or MutL defects (Table 1) [3,11,17–19]. Moreover, one isolate (CF173-1991) had a mutation in *mutY* and a molecular signature consistent with a MutY defect (*i.e.* a high proportion of transversions) (Table 1).

We did not find other mutations in *mutS* or *mutL* among the remaining genome sequences, but three additional early isolates (CF84-1972, CF43-1973, and CF105-1973) had mutations in *mutY* as well as having the molecular signature associated with a MutY defect (Table 1).

In total, we found 11 hypermutator strains among the 55 isolates. These mutators were found in ten of the 21 patients in our study (48%), which is comparable to previous findings (36%) [11]. Our results include two patients (CF211 and CF224) from whom we isolated both hypermutators and normal (normomutator) clones documenting the co-existence of both types. Indeed, the identification of both a hypermutable and a normal sub-lineage in years 1997 and 2006 from patient CF211 suggests at least 9 years of co-existence within this patient (Figure 2). It is possible that the sub-lineages with different mutation rates occupy different niches within the hosts, each niche representing different selection pressures.

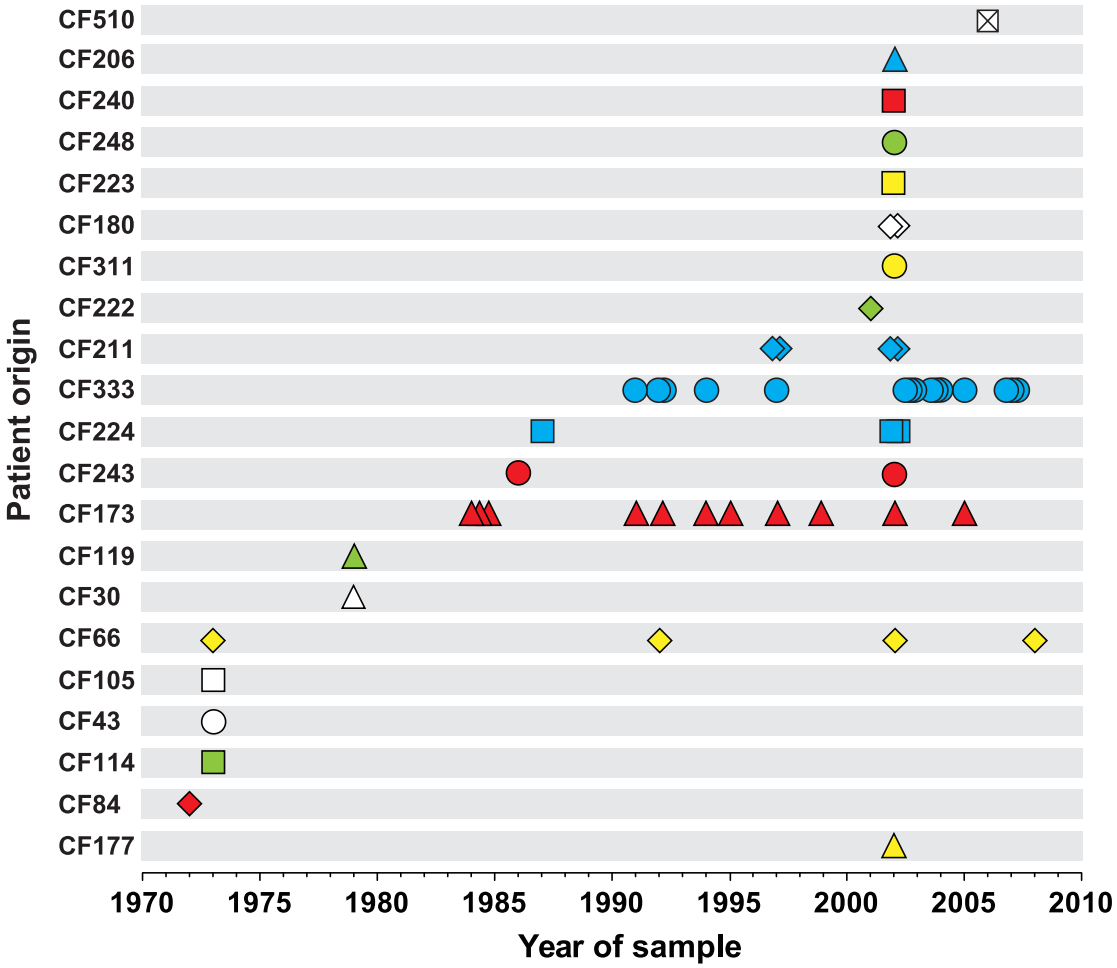


Figure 1. Patient origin and sampling time of genome sequenced *P. aeruginosa* DK2 isolates. The collection of 55 *P. aeruginosa* isolates of the DK2 clone type was sampled from 21 different CF patients over 38 years. Bacterial isolates are indicated by symbols, and if multiple isolates were sampled the same year from a patient, they are represented by stacked symbols. The isolates are named from the patient from whom they were isolated, and their isolation year (*e.g.* isolate CF173-1991).
doi:10.1371/journal.pgen.1003741.g001

Mutation rate of homopolymers

We next designed our mutational analysis to detect small insertions and deletions (microindels). A total of 1,204 unique microindels were discovered. The inheritance was explained by 1,380 parsimonious events and was congruent with the SNP-based phylogeny although the consistency for the microindels was lower (0.87) than for the SNPs (0.99). The higher rate of homoplasy among microindels would a priori be expected as microindels accumulate with high rates at mutational hotspots consisting of simple sequence repeats (SSRs) [20]. Accordingly, 93% of the inconsistent microindels were located in SSRs. As expected from current knowledge [21,22], we observed that the seven *mutL/mutS* hypermutators were particularly prone to mutation within SSRs consisting of homopolymers, and as a result 86% of the microindels that accumulated in the *mutS/mutL* hypermutable sub-lineages were localized in homopolymers whereas this was only true for 21% of the microindels within the remaining sub-lineages (Table 1).

Highly mutable loci have been shown to be important for pathogenesis and host adaptation of several pathogens [23]. For example, increased mutation rates of homopolymers in MMR-deficient *P. aeruginosa* strains have been shown *in vitro* to be important for mutational inactivation of the regulatory gene *mucA* [24], which is pivotal for adaptation in CF airways. Nonetheless, we only have a limited understanding of the homopolymer mutation dynamics at a genome-wide level and of the impact of increased mutation rates of homopolymers in relation to host adaptation. However, our collection of genome sequences from both normal and hypermutator isolates, sampled from the airways of CF patients, provides an opportunity to shed new light on homopolymer mutation rates and their impact on adaptation.

For each of the seven *mutS/mutL* hypermutator sub-lineages we calculated the mutation rates of homopolymers of different lengths (Figure 4). We observed that longer homopolymers were more likely mutated than homopolymers of shorter lengths, and for homopolymers of 3–6 nucleotides length the mutation rate increased exponentially ($R = 0.995$; Student's *t*-test, $P = 0.0026$). One might expect large homopolymers to exhibit higher probabilities of mutation, because they are distributed more frequently outside coding regions. However, we observed no evidence of this playing a role, as mutation rates of intergenic and intragenic homopolymers were similar (Figure S2). Instead, the size-dependent mutation rate of homopolymers is likely to be a

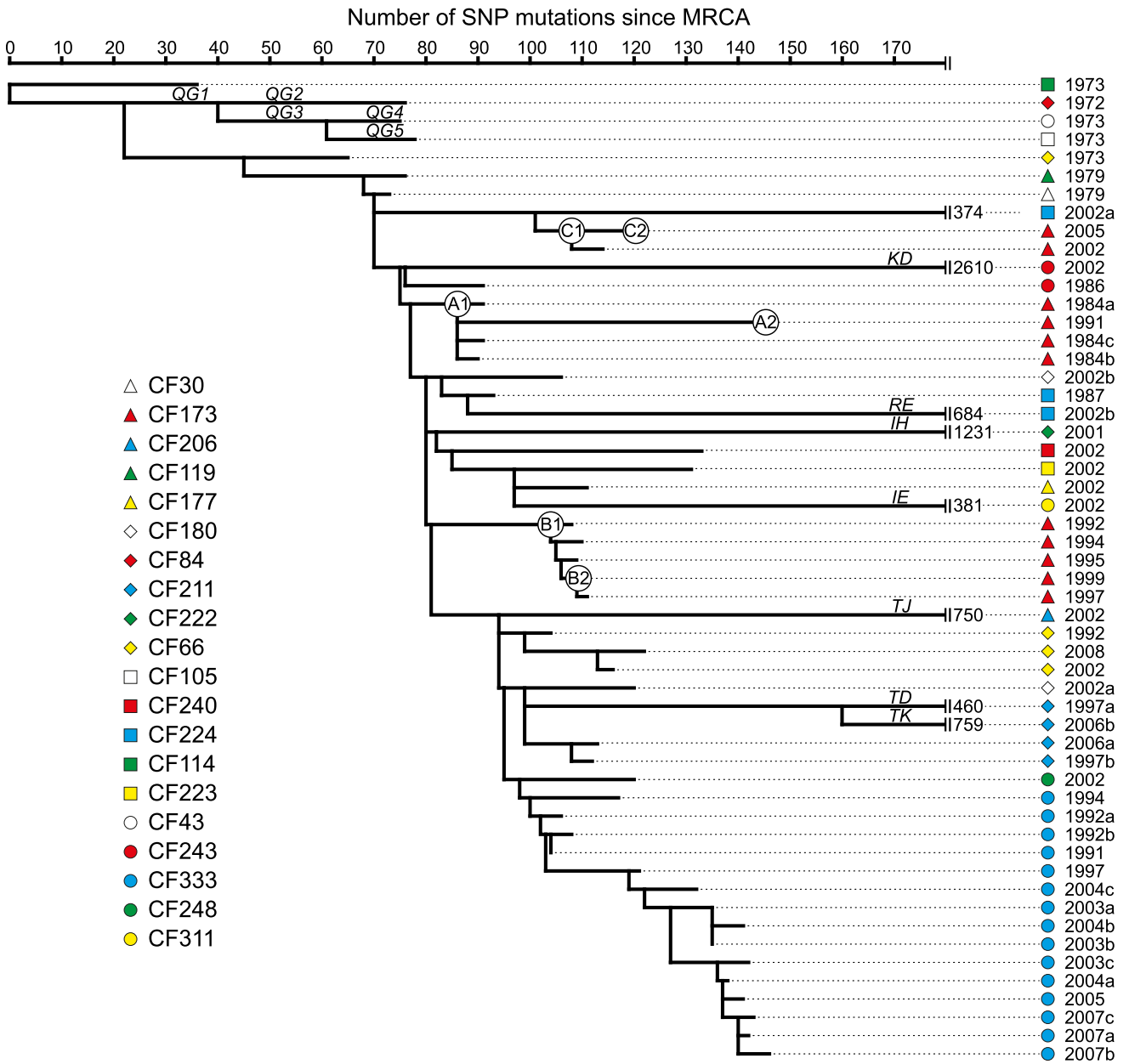


Figure 2. Maximum-parsimonious reconstruction of the phylogeny of the *P. aeruginosa* DK2 clones. The phylogenetic tree is based on 7,326 SNPs identified from whole-genome sequencing, and lengths of branches are proportional to the number of mutations. Outlier isolate CF510-2006 [9] (not shown) was used as an outgroup to determine the root of the tree. Branches leading into *mutS*, *mutL*, or, *mutY* hypermutable isolates are named as indicated by italic letters. Statistics on mutations accumulated in the specific branches are summarized in Table 1. Circles labeled A1, A2, B1, B2, C1, and C2, respectively, denotes the position of the first and last genotype of each of the DK2 sub-lineages A, B, and C which were observed to have infected patient CF173.
doi:10.1371/journal.pgen.1003741.g002

consequence of the mechanistically determined probabilities of strand-slippage during replication [20].

Distinct evolutionary paths in hypermutators

The size-dependent mutation rates of homopolymers of different lengths suggest that different genes have different probabilities of mutation. In this way, certain genes, in which variation is appreciated, may harbor sequences that are more frequently mutated in contrast to essential genes in which genetic changes are strongly selected against [23]. In agreement with this, we find that genes annotated as essential genes in *P. aeruginosa* PAO1 [25] are less likely to contain large homopolymers ≥7 nt (Fisher's exact test, $P = 0.037$) (Table S1).

Survival of bacteria in human hosts has previously been suggested to be positively influenced by rapid modulation of the cell envelope. In agreement with this (and in opposition to essential genes), we find that genes functionally related with the composition of the cell envelope are more likely to contain large ≥7 nt homopolymers (Fisher's exact test, $P = 0.002$) (Table S1). This leads us to speculate that hypermutators have a selective advantage over their normal counterparts, not only because they can speed up evolution, but also because they are creating a bias

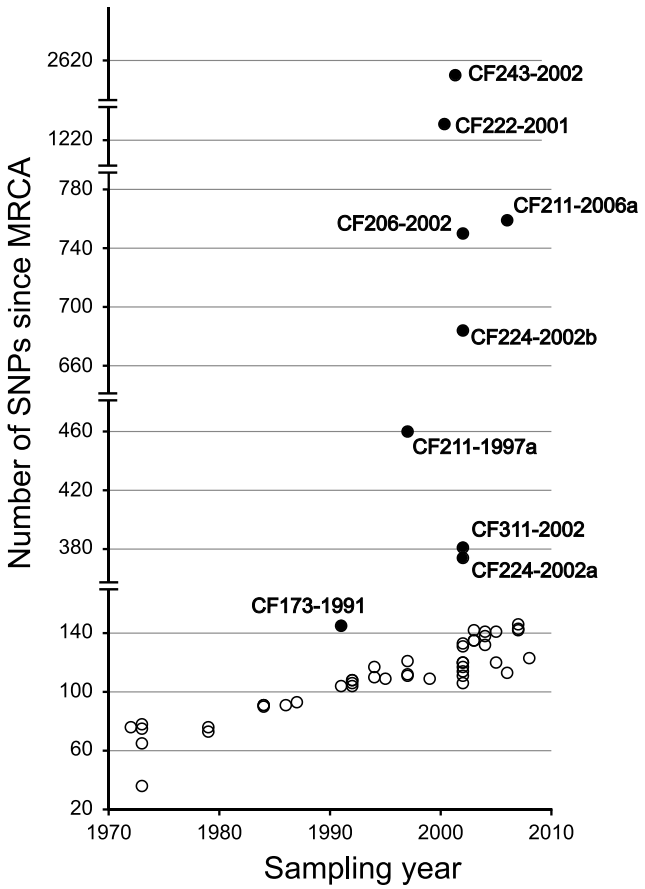


Figure 3. Total number of SNPs accumulated in each DK2 isolate. The number of SNPs accumulated in each of the isolates since their most recent common ancestor (MRCA) is plotted against isolate sampling year.
doi:10.1371/journal.pgen.1003741.g003

towards a different evolutionary path by homopolymer facilitated differential mutagenesis.

In support of our hypothesis, we find that *mutS/mutL* hypermutators acquire 3.7 more mutations in cell envelope genes containing large homopolymers (≥7 nt) relative to cell envelope genes without large homopolymers, and that the accumulation of mutations in the homopolymer-containing cell envelope genes is due to mutations within the homopolymers. Accordingly, 50% of mutations in homopolymer-containing cell envelope genes are indels whereas this is only true for 5% of the mutations in the remaining genes (9/164 vs. 8/8; Fisher's exact test, $P = 6.2 \times 10^{-6}$).

In further support of our hypothesis on differential mutagenesis we find two genes (PADK2_15360 and PADK2_03970) in which all seven *mutS/mutL* hypermutators, but no other isolates, carries mutations. Given the number of mutations within each of the seven hypermutable sub-lineages and all other lineages this observation is highly unexpected by chance ($P(X \geq 2) \sim binom(X; 5976; 2.9 \times 10^{-7}) = 1.6 \times 10^{-6}$; where 2.9×10^{-7} is the probability of an average length gene to be mutated in only the *mutS/mutL* sub-lineages). One of the genes, PADK2_15360, encodes an outer membrane receptor protein, and all seven hypermutators are independently mutated in the same 7×G homopolymer located at position 1127–1133 within the 2958 nt gene. Since none of the other 48 isolates contain mutations within PADK2_15360, we suggest that mutations in this gene represent a hypermutator-specific adaptive target for rapid modulation of the cell envelope. All seven mutations are frameshift mutations causing premature stop codons resulting in truncated proteins without a putative TonB dependent receptor domain (Pfam family PF00593) located in the C-terminal part of the protein. We hypothesize that this domain is localized in the outer membrane where it, due to its potential surface-exposure, could be a target of recognition by the immune defense.

Time-measured phylogenetic reconstruction of the DK2 lineage

To further investigate the within-host evolutionary history of the DK2 lineage and to estimate the dates of divergence between DK2

Table 1. Statistics on mutations accumulated in hypermutable sub-lineages.

Branch[a]	Clone at tip of branch	*mutS, mutL,* and, *mutY* mutations[b]		SNPs	Transitions: Transversion ratio	Microindels (% within homopolymers)
TJ	CF206-2002	*mutL(1248–1249insGCGCC)*		656	54:1	142 (90%)
TD	CF211-1997a	*mutL(C1420T)*[c]	MutL(R474W)	300	59:1	79 (94%)
TK	CF211-2006a	*mutL(C1420T)*[c]	MutL(R474W)	599	53:1	114 (82%)
IH	CF222-2001	*mutS(T347C; 1096–1097insC; G1561A)*	MutS(V116A; A521T)	1149	43:1	219 (90%)
RE	CF224-2002b	*mutL(792–802ΔTATGGTGCGCG; C1477T)*	MutS(P493S)	596	65:1	126 (86%)
KD	CF243-2002	*mutS(G506A; G1300A; C1495A), mutL(T875C; G1452A)*	MutS(R169H; E434K; R499S), MutL(V292A; M484I)	2534	86:1	268 (75%)
IE	CF311-2002	*mutS(G1567A)*	MutS(E523K)	284	40:1	104 (94%)
QG1-5	CF84-1972, CF43-1973, CF105-1973	*mutY(G925A)*	MutY(V309M)	106	1:5	11 (9%)
JC	CF173-1991	*mutY(T785A)*	MutY(L262H)	59	1:6	5 (20%)
	Other	None		1085	2:1	312 (21%)

[a]Branch names are denoted in phylogenetic tree in Figure 2.
[b]Position of mutations (consequences at the amino acid level is only shown for SNP mutations).
[c]Mutation accumulated in an ancestor shared by clones CF211-1997a and CF211-2006a.
doi:10.1371/journal.pgen.1003741.t001

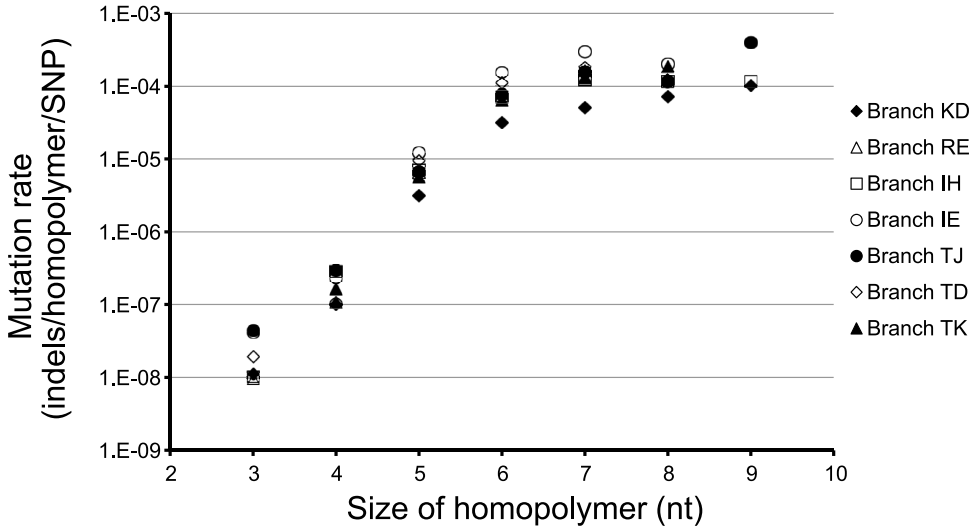


Figure 4. Mutation rates of homopolymers. Rates of mutation of homopolymers of different sizes are shown for seven DK2 sub-lineages evolving with a *mutS/mutL* DNA MMR-deficiency. The rates were calculated as the number of observed indels per homopolymer per *mutS/mutL* MMR-deficiency caused SNP (see Materials and Methods).
doi:10.1371/journal.pgen.1003741.g004

isolates, we applied Bayesian statistics to infer time-measured phylogenies using a relaxed molecular clock rate model (Figure 5).

We excluded the hypermutator isolates and isolate CF224-2002a containing recombined regions from the analysis, as they would otherwise interfere with the phylogenetic analysis. Based on this analysis, the mean mutation rate was estimated to be 2.6 SNPs/year (95% highest posterior density (HPD; see Materials and Methods) 1.8–3.2 SNPs/year) which is equivalent to 4×10^{-7} SNPs/year per site or 9×10^{-11}–11×10^{-11} SNPs/bp per generation assuming 3700–4500 generations per year [26]. Our estimated mutation rate is in the same range as those estimated for *Shigella sonnei* (6×10^{-7} SNPs/bp/year) [7] and *Vibrio cholerae* (8×10^{-7}) [2] but in between the rates reported for *Yersinia pestis* (2×10^{-8}) [6] and *Staphylococcus aureus* (3×10^{-6}) [27].

The topologies of the Bayesian phylogenetic reconstruction and the maximum-parsimonious phylogeny were congruent, and the relationship among the clones correlated with patient origin and the time of sampling (Figure 2; Figure 5).

We have previously shown that a set of specific mutations first observed in CF30-1979 and in all isolates sampled after 1979 were important for the reproductive success of the DK2 lineage and its dissemination among multiple individuals [10]. Using the phylogenetic reconstruction, we estimate that isolates sampled after 1979 diverged from a common ancestor in year 1970 (95% HPD, 1961–1976) [10]. Furthermore, our phylogenetic data document that the transmission potential of the DK2 lineage has been maintained over several decades. The most recent transmission event is predicted to have occurred in year 1997 (95% HPD, 1991–2001), as this is the latest time estimate of a predicted ancestor shared by isolates from different patients (CF177-2002 and CF223-2002). Since we have not investigated DK2 isolates from all patients chronically infected with this lineage it remains a possibility that transmission has occurred subsequent to this time.

Seven patients are represented by multiple isolates, and in six of the patients at least two of the isolates clustered as monophyletic groups according to patient origin (Figure 2; Figure 5). This is in agreement with a model in which independent sub-lineages of the DK2 clone evolved separately within individual patients, and it excludes the possibility of continuous and near-perfect mixing of strains between patients. The patient-linkage was most prevalent for patient CF333 from which all 15 isolates constituted a single monophyletic group, and the isolates branched in general according to their sampling year giving a linear evolutionary trajectory with an average distance of 6.1 SNPs (~2.3 years) from the line of descent (Figure 2; Figure 5).

In contrast, we observed an unexpected DK2 population dynamics in patient CF173 in which the isolates clustered as three different monophyletic groups with four, five and two isolates, respectively (Figure 2). This shows that patient CF173 was infected by three distinct sub-lineages rather than only a single sub-lineage. Interestingly, the three sub-lineages carried by patient CF173 can be distinguished based on the sampling year of the isolates. Accordingly, the isolates from the different clusters are sampled in the time-periods 1984–1991 (cluster A), 1992–1999 (cluster B) and 2002–2005 (cluster C), respectively. This points to a replacement of the earlier sub-lineages around years 1991–1992 and 1999–2002, respectively, caused by secondary transmission events. Alternatively, it could be the result of co-existing lineages whose time-dependent sampling was caused by shifts in relative abundance or changes in sampling probability from different niches.

Parallel evolution of genes involved in host adaptation

The presence of independently evolving DK2 sub-lineages made it possible to search for recurrent patterns of mutation and to identify bacterial genes that have acquired mutations in parallel in different individuals [1,28]. Overall, we found no evidence for either intragenic bias of the mutations or for positive selection within coding regions (dN/dS = 0.66 including all mutations; Text S1), and we would therefore expect the 7,383 intragenic mutations to be distributed randomly among the 5,976 *P. aeruginosa* DK2 genes. This means that on average a gene would acquire 1.2 mutations, and we would expect only 1.3 genes to acquire mutations more than 6 times ($P(X>6)\sim binom(X;\ 7{,}383;\ 5976^{-1}) = 2.2\times10^{-4}$). Nonetheless, we identified 65 genes that were mutated more than 6 times when comparing across all DK2 sub-lineages (see Table S2 for the full list of all 65 genes). The high mutation number within these genes could be the result of a

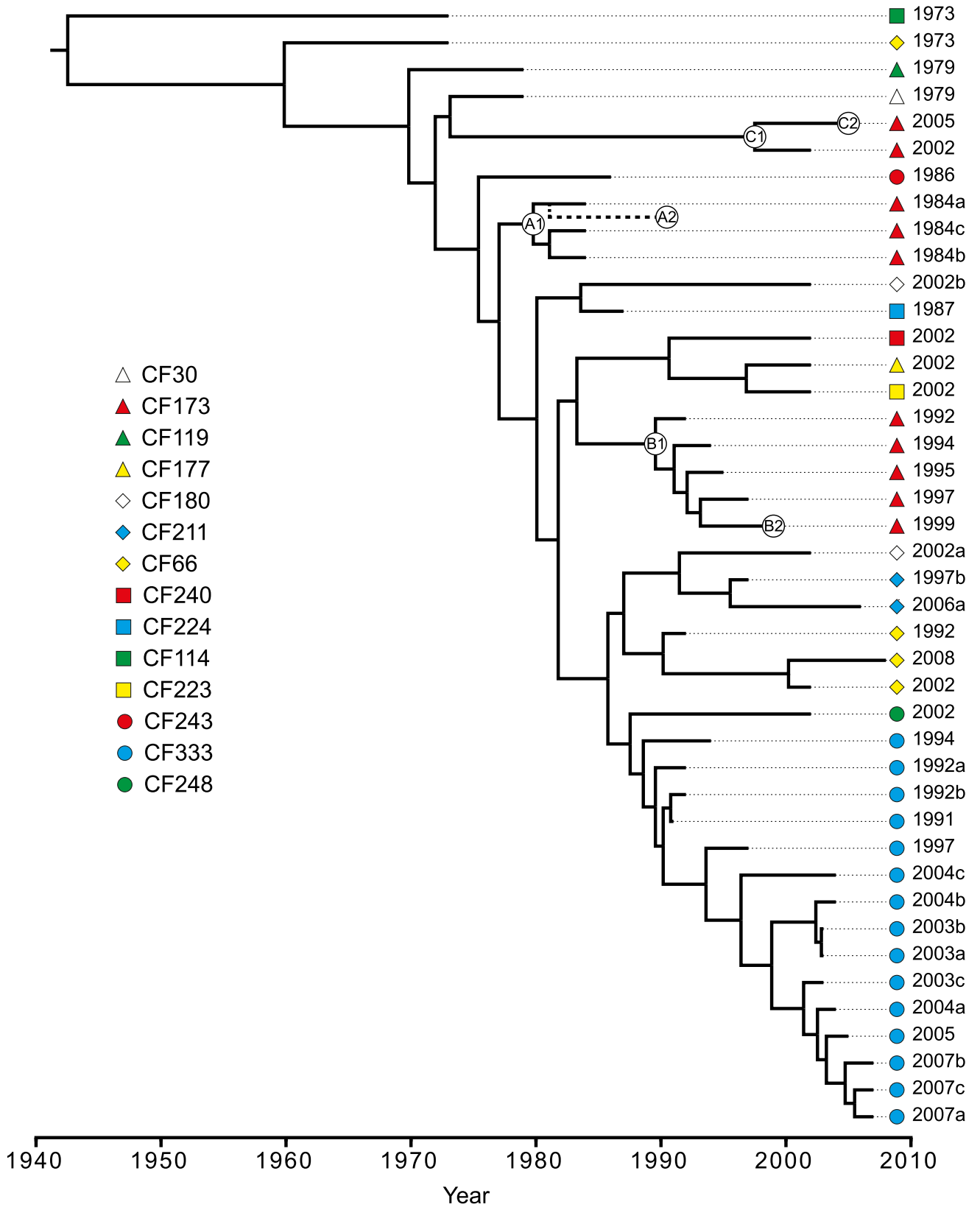


Figure 5. Bayesian phylogenetic reconstruction and divergence date estimates of the *P. aeruginosa* DK2 clones. Bayesian statistics were used to estimate the divergence times of predicted ancestors. The tree was based on 736 unique SNPs identified from whole-genome sequencing. Circles labeled A1, A2, B1, B2, C1, and C2, respectively, denotes the position of the first and last genotype of each of the DK2 sub-lineages A, B, and C which were observed to have infected patient CF173. The position of CF173-1991 (A2) is approximated from an equivalent Bayesian phylogenetic analysis including the hypermutator isolates (Figure S3).
doi:10.1371/journal.pgen.1003741.g005

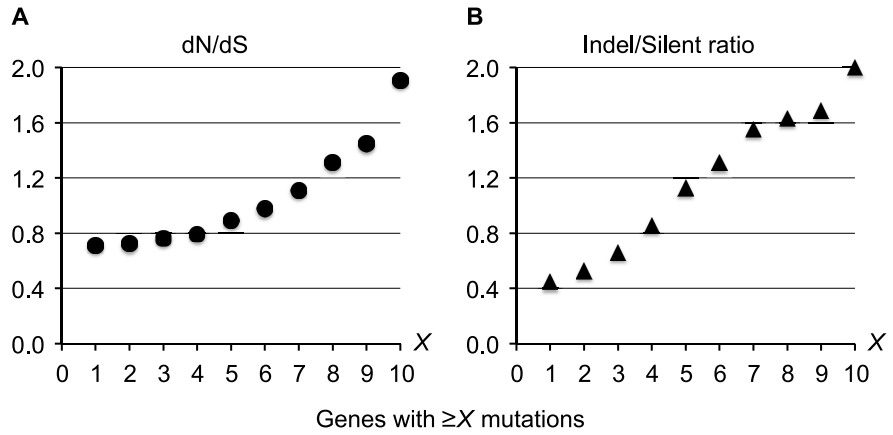


Figure 6. Increased pressures of selection for mutations in the top most mutated genes. Measures of the selection pressures were plotted for genes acquiring ≥*X* mutations during the evolution of the DK2 lineage. Plot A shows the dN/dS ratio, and plot B shows the ratio of indels relative to silent SNPs.
doi:10.1371/journal.pgen.1003741.g006

positive selection for mutations, which is supported by our observation that increased pressures of selection acts on the top most mutated genes (Figure 6). Accordingly, the signature for selection for SNPs accumulated in the 65 top most mutated genes (dN/dS = 1.11) was positive and significantly higher than for SNPs accumulated in other genes (dN/dS = 0.69; Fisher's exact test, $P = 5.2 \times 10^{-5}$).

These findings suggest that the 65 genes with multiple mutations undergo adaptive evolution (*i.e.* they are pathoadaptive genes involved in host adaptation), although the presence of neutral mutational hotspots or fast acquisition of secondary mutations within the same gene may contribute to the high mutation number in some genes. To exclude the possibility that the high mutation numbers were the result of recombination events or because of particularly large gene sizes, we left out mutations from recombined regions and large genes (>5 kb) from our analysis.

A large part of the identified pathoadaptive genes were associated with antibiotic resistance (n = 14), including the genes *ampC*, *emrB*, *ftsI*, *fusA*, *gyrA/B*, *mexB/Y*, *pmrB*, *pprA*, *oprD*, and *rpoB/C* (Figure 7 and Table S2), in which mutations have been shown to confer resistance against a range of antibiotics, *e.g.* beta-lactams, tetracyclines, quinolones, chloramphenicol, macrolides, fusidic acid, aminoglycosides, polymycins and penicillins [29–37]. As such, the detection of multiple mutations in known antibiotic resistance genes confirmed the ability of our approach to identify genes involved in host adaptation. The exact amino acid changes caused by nine out of 16 unique non-synonymous mutations found within the genes *gyrA/B* and *rpoB* have previously been shown to confer resistance against fluoroquinolones and rifampicin, respectively (Table S3).

Another major group of pathoadaptive genes (n = 18) were functionally related to the cell envelope (Figure 7 and Table S2). Possibly, these mutations have been selected to evade the host immune response [38] or, especially in the case of *lpxO2*, to prevent interaction from LPS-targeting antibiotics [39]. Also, mutations in 13 genes involved in gene regulation were identified in our analysis, suggesting that remodeling of regulatory networks is a key evolutionary pathway in host adaptation as it seems to be in evolving *Escherichia coli* populations [40]. Among the regulatory genes that acquired mutations were four yet uncharacterized genes encoding components of two-component regulatory systems, a gene-category which is significantly overrepresented (88/5823 vs. 7/58 Fisher's exact test, $P = 6.7 \times 10^{-5}$) among the pathoadaptive genes (Table S2). We suggest that these uncharacterized regulatory genes as well as other genes identified as involved host adaptation represent potential therapeutic targets.

Mutations in pathoadaptive genes correlate with strain displacement events within patient CF173

The adaptive benefits of a mutation are usually investigated by introduction of single or multiple mutations into isogenic strains and testing for fitness effects associated with the mutation(s) in controlled experimental conditions (such as competition experiments). Such testing is most effective when the phenotype (*e.g.* antibiotic resistance) can be easily interpreted in relation to the fitness impact. However, for mutations for which no or only subtle phenotypic changes are apparent it is difficult to directly test the fitness effects. In addition, the impacts on fitness of specific mutations must be assessed in the same environment as the one in which the mutation was selected. This is obviously not possible in case of human airway infections. To circumvent these limitations, we hypothesize that the count of mutations within the pathoadaptive genes can be used as a measure of the fitness of individual clones of *P. aeruginosa*. To investigate this hypothesis we took advantage of the two strain displacements (or changes in strain abundances) that occurred in patient CF173 in the years 1991–1992 and 1999–2002, which suggested that CF173 was infected by three succeeding DK2 sub-lineages A (1980–1991), B (1990–1999), and C (2000–2005).

We assume that the succeeding sub-lineage must be better adapted (*i.e.* having a higher fitness) than the previous sub-lineage, which was outcompeted. When determining the number of mutations found in the sub-lineages A, B, and C within the pathoadaptive genes, it was striking that the succeeding genotype consistently had a higher count of mutations than the previous genotype (Figure 7). In this way, the counts of mutations correlated with the strain displacement observed within patient CF173. We suggest that the mutation count can be used to predict the fitness of emerging DK2 clones, and that the pathogenicity scoring together with the information about the specific mutations can be used as a novel approach for clinicians to treat and segregate patients. It should be noted that our results cannot simply be ascribed to the succeeding genotypes having more mutations in general as no significant positive correlation existed between the

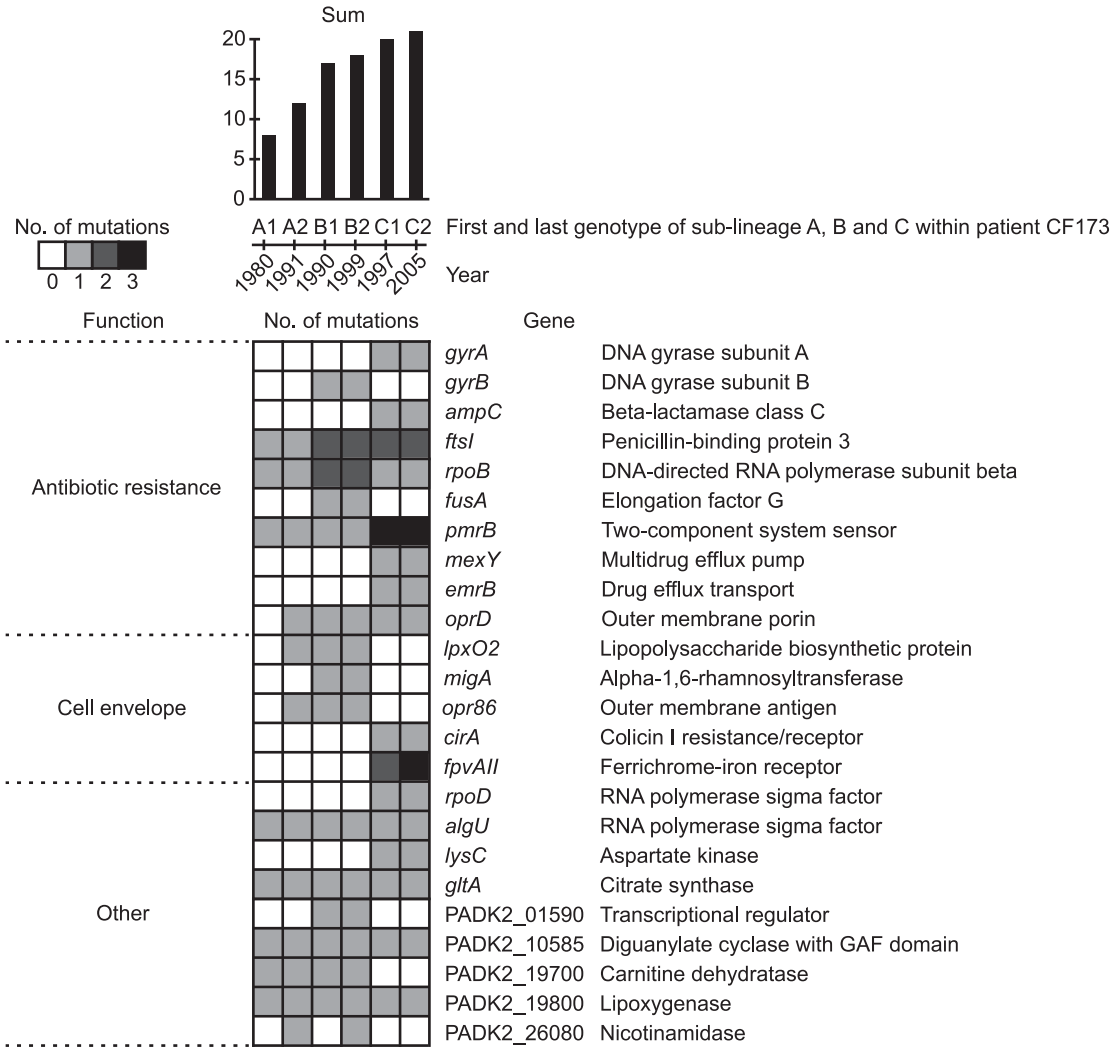


Figure 7. Pathoadaptive genes. Genes identified from parallel evolution to be involved in host adaptation. Colors of squares denotes if the gene was mutated relative to the MRCA of the DK2 clones. Only genes mutated in any of the isolates from CF173 are shown (see full list of pathoadaptive genes in Table S2). The presence of mutations is shown for the first and last genotypes of each of the DK2 sub-lineages A, B, and C, which were observed to have infected patient CF173. The total sum of mutations observed within each of the genotypes is indicated at the top. Genes are grouped by function. Details about specific mutations and their fixation in the DK2 isolates are given in Table S5, Table S6, and Figure S4.
doi:10.1371/journal.pgen.1003741.g007

total number of mutations and the number of mutations within the pathoadaptive genes ($R = 0.30$; Student's *t*-test, $P = 0.28$).

Conclusions and implications

By genome sequencing of 55 isolates of the transmissible DK2 clone type of *P. aeruginosa*, we have provided a detailed view of the evolution of a bacterial pathogen within its human host. The sampling from multiple patients offered the opportunity to detect loci that were independently mutated in parallel lineages, here referred to as pathoadaptive genes, whereas sampling multiple times from the same patient gave an opportunity to study the within-patient population dynamics.

Several of the pathoadaptive genes identified here were associated with antibiotic resistance, gene regulation, and composition of the cell envelope. Some of these genes have been found in other studies of genomic evolution in CF pathogens to be important for adaptation [1,3,17]. Genomic analysis of additional *P. aeruginosa* lineages from different patients and clinical settings will enable a systematic identification of genes that are repeated targets for selective mutations during adaptation to life in the CF lung. Importantly, we also identified genes of unknown function and without prior implication in pathogenesis. Further investigations of the function of these genes are required to determine their potential as future therapeutic targets against the infection.

An exceptional 21-year time series of 11 isolates sampled from patient CF173 revealed a complex population dynamics in which the patient was infected by three distinct sub-lineages of the DK2 clone type, each sub-lineage being dominant over several years until its final decline or disappearance. This observation illustrates the power of high-throughput sequencing in relation to uncovering pathogen dynamics within infected individuals. We further observed that the cumulative count of mutations within pathoadaptive genes increased for each of the succeeding sub-lineages. This means that emerging sub-lineages carried a cumulative palette of pathoadaptive mutations and not only adaptive mutations conferring an advantage for a newly introduced selection force that may have triggered the removal of the preceding lineage.

The identification of pathoadaptive genes involved in host adaptation and our finding that the specific count of mutations within these genes act as a classifier that predict the pathogenicity of emerging sub-lineages of the DK2 clone type, should enable better epidemiological predictions and provide valuable information for the clinicians on how to treat and segregate patients.

The presence of hypermutable lineages within 48% of the studied individuals might be the outcome of an accelerated acquisition of beneficial mutations within hypermutators [11,41–43]. Nonetheless, our examination of mutation dynamics of homopolymers provided a novel genome-wide perspective on the impact and potential advantage of differential mutagenesis associated with the hypermutator phenotype. Showing a clear exponential correlation between the rate of change and the size of the homopolymer, we confirmed homopolymers to be hotspots for differential mutagenesis, and we identified two homopolymer-containing genes to be preferentially mutated in hypermutators.

In conclusion, we have shown how collections of isolates of bacteria sampled from chronically infected patients constitute a valuable basis for studying evolution of pathogens *in vivo*, and our results facilitates comparative studies as sequencing datasets become increasingly available.

Materials and Methods

Bacterial strains and genome sequencing

The study encompasses 55 isolates of the *P. aeruginosa* DK2 clone type that were sampled over 38 years from 21 CF patients attending the Copenhagen Cystic Fibrosis Center at the University Hospital, Rigshospitalet (Figure 1). Isolation and identification of *P. aeruginosa* from sputum was done as previously described [44]. Sequencing of 45 of the isolates was previously reported by Yang *et. al.* [10] and Rau *et al.* [9]. Two of the previously sequenced isolates (CF333-1991 and CF510-2006) were re-sequenced together with ten new isolates on an Illumina HiSeq2000 platform generating 100-bp paired-end reads using a multiplexed protocol to an average coverage depth of 63–212 fold. Sequence reads from all isolates are deposited in the Short Read Archive under accession number ERP002277 (accession numbers for individual samples are provided in Table S4).

Mutation detection and analysis

Reads were mapped against the *P. aeruginosa* DK2 reference genome (CF333-2007a; Genbank accession no. CP003149) using Novoalign (Novocraft Technologies) [45], and pileups of the read alignments were produced by SAMtools release 0.1.7 [46]. Single nucleotide polymorphisms were called by the varFilter algorithm in SAMtools in which minimum SNP coverage was set to 3 (samtools.pl varFilter -d 3 -D 10000). Only SNP calls with quality scores (Phred-scaled probability of sample reads being homozygous reference) of at least 50 (*i.e.* $P \leq 10^{-5}$) were retained. Microindels were extracted from the read pileup by the following criteria; (1) quality scores of at least 500, (2) root-mean-square (RMS) mapping qualities of at least 25, and (3) support from at least one fifth of the covering reads. The false-negative rates were found to be 2% and 3% by *in silico* introduction of random base-substitutions and microindels (lengths 1–10 bp), respectively. To avoid false-positives, the reference genome was re-sequenced by Illumina sequencing to exclude polymorphisms caused by errors in reference assembly. Also, Illumina re-sequencing of CF333-1991 confirmed all the SNPs (and found no other SNPs) that were previously reported for this isolate by use of pyrosequencing [10]. Indeed, the confirmation by re-sequencing of CF333-1991 and the fact that many isolates are only discriminated by a few mutations verify that our genomic analysis has a very low false-positive rate.

A maximum-parsimonious phylogenetic analysis was used to predict the relationship and mutational events among the clones of the DK2 clone type. The tree consistency index ($CI = m/s$) was calculated as the minimum number of changes (m) divided by the number of changes required on the tree (s). The CI will equal 1 when there is no homoplasy.

For the calculation of average distances of the 15 CF333 isolates to their line of descent, the line of descent was defined as the direct lineage from the most recent common ancestor (MRCA) of all 15 isolates until the MRCA of the three most recently sampled isolates (CF333-2007a, CF333-2007b, CF333-2007c).

To provide the most accurate estimates of the relative homopolymer mutation rates in the *mutS/mutL* MMR-deficient sub-lineages, we calculated the rates per *mutS/mutL* MMR-deficiency caused SNP. This corrected count of SNPs were found by subtracting the fraction of SNPs expected to have accumulated due to the normal underlying mutation rate, *i.e.* SNPs not caused by the *mutS/mutL* MMR-deficiency. For this purpose a 2:1 transition to transversion ratio was assumed for the normal background mutation rate. This means that the SNP count of hypermutator branch "KD" composed of 2,534 SNPs (Table 1), hereof 29 transversions, was corrected to 2,447 *mutS/mutL* MMR-deficiency caused SNPs. All results and conclusions were unaffected from this correction.

Bayesian evolutionary analysis

Bayesian analysis of evolutionary rates and divergence times was performed using BEAST v1.7.2 [47]. BEAST was run with isolate CF510-2006 as an outgroup [9] and the following user-determined settings; a lognormal relaxed molecular clock model which allows rates of evolution to vary amongst the branches of the tree, and a general time-reversible substitution model with gamma correction. Results were produced from three independent chains of 50 million steps each, sampled every 5,000 steps. The first 5 million steps of each chain were discarded as a burn-in. The results were combined, and the maximum clade credibility tree was generated (using LogCombiner and TreeAnnotator programs from the BEAST package, respectively). The effective sample-sizes (ESS) of all parameters were >500 as calculated by Tracer v1.5 (available from http://beast.bio.ed.ac.uk/Tracer), which was also used to calculate 95% HPD confidence intervals of the mutation rate (*i.e.* an interval in which the modeled parameter resides with 95% probability). The root of the tree was predicted to be in year 1943 (95% HPD, 1910–1962). Note, as this estimate is based on isolates primarily sampled after year 1980, the same accumulation rate of SNPs might not hold true for the evolution of the DK2 clone type before 1980.

Supporting Information

Figure S1 Distribution of SNPs accumulated in the branch leading to CF224-2002a. Genomic overview of the distribution of SNPs accumulated in the branch leading to isolate CF224-2002a according to the predicted phylogeny in Figure 2. The positions of the two genomic regions, in which the majority of the SNPs were found, and the remaining 8 SNPs are all indicated by text. The results provide evidence of two events in which imported DNA have recombined into the chromosome of CF224-2002a. The high density of polymorphisms suggests the imported DNA to origin from a *P. aeruginosa* strain(s) unrelated to the DK2 clone type. No

genes related to mobilization and transfer of DNA was present within a 10 kb range of the predicted boundaries of the horizontally transferred regions, so we have no evidence for elements specialized for horizontal gene transfer to play a role.
(EPS)

Figure S2 Mutation rates of homopolymers in intragenic and intergenic regions. Rates of mutation of homopolymers of different sizes are shown for intragenic and intergenic homopolymers, respectively. The rates are averages of the seven DK2 sub-lineages evolving with a *mutS/mutL* MMR-deficiency. The rates were calculated as the number of observed indels per homopolymer per *mutS/mutL* MMR-deficiency caused SNP (see Materials and Methods).
(EPS)

Figure S3 Bayesian phylogenetic reconstruction and divergence date estimates of the *P. aeruginosa* DK2 clones. Bayesian statistics were used to estimate the divergence times of predicted ancestors. The tree is based on 7,326 unique SNPs identified from whole-genome sequencing. Circles labeled A1, A2, B1, B2, C1, and C2, respectively, denotes the position of the first and last genotype of each of the DK2 sub-lineages A, B, and C which were observed to have infected patient CF173.
(EPS)

Figure S4 Maximum-parsimonious reconstruction of the phylogeny of the *P. aeruginoisa* DK2 clones. The phylogenetic tree is based on 8,530 mutations (SNPs and indels) identified from whole-genome sequencing. Outlier isolate CF510-2006 (not shown) was used as an outgroup to root the tree. Italic letters indicate branch names, and lengths of branches are proportional to the number of mutations. The specific mutations that have accumulated during each specific branch are listed in Table S5 (SNPs) and Table S6 (indels).
(EPS)

Table S1 Prevalence of homopolymers in essential genes and genes functionally related to the composition of the cell envelope.
(DOCX)

Table S2 Full list of pathoadaptive genes (n = 65).
(XLSX)

Table S3 Non-synonymous mutations in genes *rpoB*, *gyrA* and *gyrB*. For each of the mutations, we have listed studies of our knowledge to have shown or indicated the specific mutation to confer resistance against an antibiotic.
(DOCX)

Table S4 List of sequenced samples and corresponding Sequence Read Archive (SRA) accession numbers.
(XLSX)

Table S5 Full list of SNPs.
(XLSX)

Table S6 Full list of indels.
(XLSX)

Text S1 Mutational signature of genetic drift.
(DOCX)

Acknowledgments

We thank Trine Markussen and Søren Damkiær for assistance with genomic DNA library preparation; Morten O.A. Sommer for helpful suggestions and expert assistance with genome sequencing; Niels Høiby for clinical information; and Henrik Aanæs and Martin S. Christiansen for helpful discussion on the hypermutator mutation rates.

Author Contributions

Conceived and designed the experiments: RLM SM LJ. Performed the experiments: RLM. Analyzed the data: RLM SM LJ. Contributed reagents/materials/analysis tools: HKJ. Wrote the paper: RLM SM LJ.

References

1. Lieberman TD, Michel JB, Aingaran M, Potter-Bynoe G, Roux D, et al. (2011) Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. Nat Genet 43: 1275–1280.
2. Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, et al. (2011) Evidence for several waves of global transmission in the seventh cholera pandemic. Nature 477: 462–465.
3. Smith EE, Buckley DG, Wu Z, Saenphimmachak C, Hoffman LR, et al. (2006) Genetic adaptation by Pseudomonas aeruginosa to the airways of cystic fibrosis patients. Proc Natl Acad Sci U S A 103: 8487–8492.
4. Harris SR, Clarke IN, Seth-Smith HM, Solomon AW, Cutcliffe LT, et al. (2012) Whole-genome analysis of diverse Chlamydia trachomatis strains identifies phylogenetic relationships masked by current clinical typing. Nat Genet 44: 413–419, S411.
5. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, et al. (2011) Rapid pneumococcal evolution in response to clinical interventions. Science 331: 430–434.
6. Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, et al. (2010) Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity. Nat Genet 42: 1140–1143.
7. Holt KE, Baker S, Weill FX, Holmes EC, Kitchen A, et al. (2012) Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. Nat Genet 44: 1056–1059.
8. Sokurenko EV, Hasty DL, Dykhuizen DE (1999) Pathoadaptive mutations: gene loss and variation in bacterial pathogens. Trends Microbiol 7: 191–195.
9. Rau MH, Marvig RL, Ehrlich GD, Molin S, Jelsbak L (2012) Deletion and acquisition of genomic content during early stage adaptation of Pseudomonas aeruginosa to a human host environment. Environ Microbiol 14: 2200–2211.
10. Yang L, Jelsbak L, Marvig RL, Damkiaer S, Workman CT, et al. (2011) Evolutionary dynamics of bacteria in a human host environment. Proc Natl Acad Sci U S A 108: 7481–7486.
11. Oliver A, Canton R, Campo P, Baquero F, Blazquez J (2000) High frequency of hypermutable Pseudomonas aeruginosa in cystic fibrosis lung infection. Science 288: 1251–1254.
12. Matic I, Radman M, Taddei F, Picard B, Doit C, et al. (1997) Highly variable mutation rates in commensal and pathogenic Escherichia coli. Science 277: 1833–1834.
13. Mao EF, Lane L, Lee J, Miller JH (1997) Proliferation of mutators in A cell population. J Bacteriol 179: 417–422.
14. LeClerc JE, Li B, Payne WL, Cebula TA (1996) High mutation frequencies among Escherichia coli and Salmonella pathogens. Science 274: 1208–1211.
15. Richardson AR, Yu Z, Popovic T, Stojiljkovic I (2002) Mutator clones of Neisseria meningitidis in epidemic serogroup A disease. Proc Natl Acad Sci U S A 99: 6103–6107.
16. Sniegowski PD, Gerrish PJ, Lenski RE (1997) Evolution of high mutation rates in experimental populations of E. coli. Nature 387: 703–705.
17. Chung JC, Becq J, Fraser L, Schulz-Trieglaff O, Bond NJ, et al. (2012) Genomic Variation among Contemporary Pseudomonas aeruginosa Isolates from Chronically Infected Cystic Fibrosis Patients. J Bacteriol 194: 4857–4866.
18. Oliver A, Baquero F, Blazquez J (2002) The mismatch repair system (mutS, mutL and uvrD genes) in Pseudomonas aeruginosa: molecular characterization of naturally occurring mutants. Mol Microbiol 43: 1641–1650.
19. Schaaper RM, Dunn RL (1987) Spectra of spontaneous mutations in Escherichia coli strains defective in mismatch correction: the nature of in vivo DNA replication errors. Proc Natl Acad Sci U S A 84: 6220–6224.
20. Bichara M, Wagner J, Lambert IB (2006) Mechanisms of tandem repeat instability in bacteria. Mutat Res 598: 144–163.
21. Strand M, Prolla TA, Liskay RM, Petes TD (1993) Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature 365: 274–276.
22. Levinson G, Gutman GA (1987) High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in Escherichia coli K-12. Nucleic Acids Res 15: 5323–5338.
23. Moxon ER, Rainey PB, Nowak MA, Lenski RE (1994) Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr Biol 4: 24–33.
24. Moyano AJ, Smania AM (2009) Simple sequence repeats and mucoid conversion: biased mucA mutagenesis in mismatch repair-deficient Pseudomonas aeruginosa. PLoS One 4: e8203.
25. Liberati NT, Urbach JM, Miyata S, Lee DG, Drenkard E, et al. (2006) An ordered, nonredundant library of Pseudomonas aeruginosa strain PA14 transposon insertion mutants. Proc Natl Acad Sci U S A 103: 2833–2838.

26. Yang L, Haagensen JA, Jelsbak L, Johansen HK, Sternberg C, et al. (2008) In situ growth rates and biofilm development of Pseudomonas aeruginosa populations in chronic lung infections. J Bacteriol 190: 2767–2776.
27. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, et al. (2010) Evolution of MRSA during hospital transmission and intercontinental spread. Science 327: 469–474.
28. Marvig RL, Sondergaard MS, Damkiaer S, Hoiby N, Johansen HK, et al. (2012) Mutations in 23S rRNA confer resistance against azithromycin in Pseudomonas aeruginosa. Antimicrob Agents Chemother 56: 4519.
29. Severinov K, Markov D, Severinova E, Nikiforov V, Landick R, et al. (1995) Streptolydigin-resistant mutants in an evolutionarily conserved region of the beta' subunit of Escherichia coli RNA polymerase. J Biol Chem 270: 23926–23929.
30. Yee YC, Kisslinger B, Yu VL, Jin DJ (1996) A mechanism of rifamycin inhibition and resistance in Pseudomonas aeruginosa. J Antimicrob Chemother 38: 133–137.
31. Liao X, Hancock RE (1995) Cloning and characterization of the Pseudomonas aeruginosa pbpB gene encoding penicillin-binding protein 3. Antimicrob Agents Chemother 39: 1871–1874.
32. Lomovskaya O, Lewis K (1992) Emr, an Escherichia coli locus for multidrug resistance. Proc Natl Acad Sci U S A 89: 8938–8942.
33. Bielecki P, Lukat P, Husecken K, Dotsch A, Steinmetz H, et al. (2012) Mutation in elongation factor G confers resistance to the antibiotic argyrin in the opportunistic pathogen Pseudomonas aeruginosa. Chembiochem 13: 2339–2345.
34. Hancock RE (1998) Resistance mechanisms in Pseudomonas aeruginosa and other nonfermentative gram-negative bacteria. Clin Infect Dis 27 Suppl 1: S93–99.
35. Strateva T, Yordanov D (2009) Pseudomonas aeruginosa - a phenomenon of bacterial resistance. J Med Microbiol 58: 1133–1148.
36. Moskowitz SM, Ernst RK, Miller SI (2004) PmrAB, a two-component regulatory system of Pseudomonas aeruginosa that modulates resistance to cationic antimicrobial peptides and addition of aminoarabinose to lipid A. J Bacteriol 186: 575–579.
37. Wang Y, Ha U, Zeng L, Jin S (2003) Regulation of membrane permeability by a two-component regulatory system in Pseudomonas aeruginosa. Antimicrob Agents Chemother 47: 95–101.
38. Nguyen D, Singh PK (2006) Evolving stealth: genetic adaptation of Pseudomonas aeruginosa during cystic fibrosis infections. Proc Natl Acad Sci U S A 103: 8305–8306.
39. Moskowitz SM, Brannon MK, Dasgupta N, Pier M, Sgambati N, et al. (2012) PmrB mutations promote polymyxin resistance of Pseudomonas aeruginosa isolated from colistin-treated cystic fibrosis patients. Antimicrob Agents Chemother 56: 1019–1030.
40. Philippe N, Crozat E, Lenski RE, Schneider D (2007) Evolution of global regulatory networks during a long-term experiment with Escherichia coli. Bioessays 29: 846–860.
41. Waine DJ, Honeybourne D, Smith EG, Whitehouse JL, Dowson CG (2008) Association between hypermutator phenotype, clinical variables, mucoid phenotype, and antimicrobial resistance in Pseudomonas aeruginosa. J Clin Microbiol 46: 3491–3493.
42. Ferroni A, Guillemot D, Moumile K, Bernede C, Le Bourgeois M, et al. (2009) Effect of mutator P. aeruginosa on antibiotic resistance acquisition and respiratory function in cystic fibrosis. Pediatr Pulmonol 44: 820–825.
43. Ciofu O, Riis B, Pressler T, Poulsen HE, Hoiby N (2005) Occurrence of hypermutable Pseudomonas aeruginosa in cystic fibrosis patients is associated with the oxidative stress caused by chronic lung inflammation. Antimicrob Agents Chemother 49: 2276–2282.
44. Hoiby N, Frederiksen B (2000) Microbiology of cystic fibrosis. In: Hodson M, Geddes D, editors. Cystic fibrosis. 2nd ed. London, United Kingdom: Arnold. pp. 83–107.
45. Krawitz P, Rodelsperger C, Jager M, Jostins L, Bauer S, et al. (2010) Microindel detection in short-read sequence data. Bioinformatics 26: 722–729.
46. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.
47. Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29: 1969–1973.

Within-Host Evolution of *Pseudomonas aeruginosa* Reveals Adaptation Towards Iron Acquisition from Hemoglobin

Rasmus Lykke Marvig[1,*], Søren Damkiær[1,*], S.M. Hossein Khademi[1], Trine Markussen[1], Søren Molin[1], and Lars Jelsbak[1,a]

[1] Department of Systems Biology, Technical University of Denmark, 2800 Lyngby, Denmark

[a] Corresponding author: Lars Jelsbak, Department of Systems Biology, Technical University of Denmark, 2800 Lyngby, Denmark. Email: lj@bio.dtu.dk. Telephone: +45 45256129.

* These authors contributed equally to the work.

Running title: Taste for blood: Adaptation towards heme utilization

Abstract:

Pseudomonas aeruginosa airway infections are a major cause of mortality and morbidity of cystic fibrosis (CF) patients. In order to persist *P. aeruginosa* depends on acquiring iron from its host, and multiple different iron acquisition systems may be active during infection. This includes the pyoverdine siderophore and the *Pseudomonas* heme utilization (*phu*) system. While the regulation and mechanisms of several iron-scavenging systems are well described, it is not clear whether such systems are target for selection during adaptation of *P. aeruginosa* to the host environment. Here we investigate the within-host evolution of the transmissible *P. aeruginosa* DK2 lineage. We find a positive selection for promoter mutations leading to the increased expression of *phuR* encoding an outer membrane receptor of the *phu* system. By mimicking conditions of the CF airways *in vitro*, we experimentally demonstrate that increased expression of *phuR* confers a growth advantage in the presence of hemoglobin, thus suggesting that *P. aeruginosa* genetically adapt towards iron acquisition from hemoglobin. To rule out that this adaptive trait is specific to the DK2 lineage, we inspected the genomes of two additional *P. aeruginosa* lineages isolated from CF airways, and we found similar adaptive evolution in both of these lineages. Furthermore, in all three lineages *phuR* promoter mutations coincided with the loss of pyoverdine production, suggesting that within-host adaptation towards heme utilization is triggered by the loss of pyoverdine production. Targeting heme utilization might therefore be a promising strategy for the treatment of *P. aeruginosa* infections in CF patients.

Introduction:

Iron is an essential component for virtually all forms of life. This includes bacterial pathogens that depend on acquiring iron from their hosts in order to replicate and cause disease [1]. A general defensive mechanism of the host is therefore to withhold iron from invading bacteria to prevent their growth, but this defense is counter measured by bacterial pathogens as they posses specific systems to scavenge iron from their hosts. While the regulation and mechanisms of several of such iron-scavenging systems are well described [1], not much is known about how the within-host selection pressures act on the pathogens ability to acquire iron. This is especially relevant during long-term chronic infections in which invading bacteria genetically adapt to the host environment.

The opportunistic pathogen *Pseudomonas aeruginosa* is a common environmental inhabitant which is capable of causing long-term chronic infections in the airways in patients with cystic fibrosis (CF), and *P. aeruginosa* infections are directly associated with the morbidity and mortality of CF patients. Chronic infections in CF patients provide an opportunity for long-term monitoring of the battle between the infecting bacteria and the host [2-6], and thus offer a direct method for observing the genetic adaptation of *P. aeruginosa* to the human host environment.

Most iron in the human body is bound in hemoglobin, which is an oxygen transport protein in red blood cells [1]. If not bound by essential proteins such as hemoglobin, iron is withheld and stored by binding to proteins like transferrin, lactoferrin, and ferritin. *P. aeruginosa* is known to scavenge iron from the human host by both siderophore-based systems or heme acquisition systems [7].

Siderophores are low-molecular weight molecules secreted from bacteria. The strong association of iron to siderophores enables them to remove iron from the human iron-

storage proteins whereupon the siderophore-iron complex can be taking up by cognate receptors at the bacterial surface. The major siderophore secreted by *P. aeruginosa* is pyoverdine [7], and iron-loaded pyoverdine is taken up by the outer membrane receptor FpvA [8, 9].

Alternatively, iron contained in the heme group of hemoglobin can be taken up by either of two heme uptake systems in *P. aeruginosa*. The two systems are the *Pseudomonas* heme utilization (*phu*) system and the heme assimilation system (*has*) [10]. The two systems are different in the sense that the *phu* system is dependent on the direct uptake of heme by the outer membrane receptor PhuR, whereas the *has* system encodes a secreted hemophore HasA that returns heme to an outer membrane receptor HasR.

It is not clear to what extent the different iron uptake systems in *P. aeruginosa* play a role for survival in the lungs of CF patients. Detection of pyoverdine in the sputa of some CF patients have proposed pyoverdine to play a key role in infection [11, 12]; however, quantification of transcription of iron uptake systems in sputum samples suggested multiple systems to be active and that siderophore-mediated uptake may not be dominant in all patients [13, 14].

In an effort to understand the genetic adaptation of *P. aeruginosa* to the CF airways, we recently mapped all mutational changes in the *P. aeruginosa* DK2 lineage as it spread among 21 Danish CF patients by interpatient transmission [2]. The study showed that the selective forces driving the evolution of *P. aeruginosa* in the CF airways could be inferred from convergent evolution of DK2 sub-lineages evolving in parallel in separate hosts. Here, we further analyze the genomic data and provide evidence that within-host evolution of *P. aeruginosa* is characterized by adaptation towards iron acquisition from hemoglobin.

89

90 **Results and discussion:**

91 *Parallel evolution of mutations in the phuR promoter region*

92 It is known that *P. aeruginosa* genetically adapt to CF patients during long-term
93 chronic infections, and several studies have genome sequenced isolates of *P.*
94 *aeruginosa* sampled longitudinally from the airways of CF patients to map the
95 mutations that accumulate during infection [2-6]. In one such study, we mapped all
96 the mutations that had occurred in the *P. aeruginosa* DK2 lineage during 36 years of
97 infection [2]. Whole-genome analysis of 55 DK2 isolates enabled a fine-grained
98 reconstruction of the evolutionary relationship of the DK2 lineage, and the study
99 identified several genes to be targeted by mutation to optimize pathogen fitness within
100 the host environment (pathoadaptation). Nonetheless, only intragenic mutations (*i.e.*
101 mutations within genes) were included to identify such pathoadaptive patterns of
102 mutation. Here, we therefore re-analyzed the data with respect to intergenic regions as
103 selection might act also on such sequences due to their role in regulation and
104 transcription of neighboring genes.
105 The 6,402,658 bp genome of the *P. aeruginosa* DK2 strain contains 4,883 intergenic
106 regions with an average size of 146 bp, and the intergenic regions constitute a total of
107 714,368 bp. Marvig *et al.* (2013) found 1,365 intergenic mutations meaning that one
108 would expect an average length intergenic region to be hit by 0.3 mutations (or
109 0.0019 mutation/bp). Searching for recurrent patterns of mutation of the same genetic
110 loci makes it possible to identify positive selection for mutations effecting genes
111 important for host-adaptation [2, 15, 16]. We therefore focused on the intergenic
112 regions with the highest densities of mutations, and interestingly found the 180-bp
113 intergenic region containing the promoters of the *phu* system to be the most

frequently mutated with a total of 13 mutations (0.072 mutation/bp) (Figure 1). This number of mutations is 38-fold higher than what would be expected by chance and represents a significant increase in mutation density ($P(X \geq 0.072) \sim pois(X; 0.0019) = 0.002$).

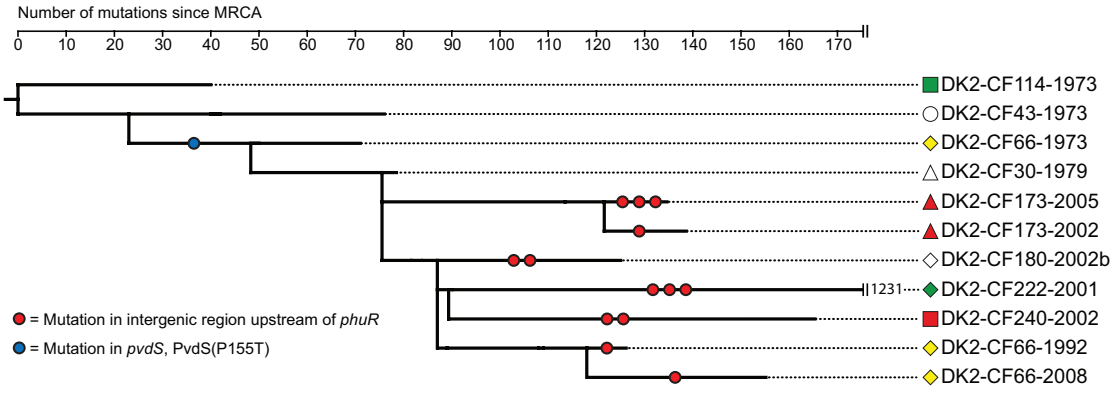
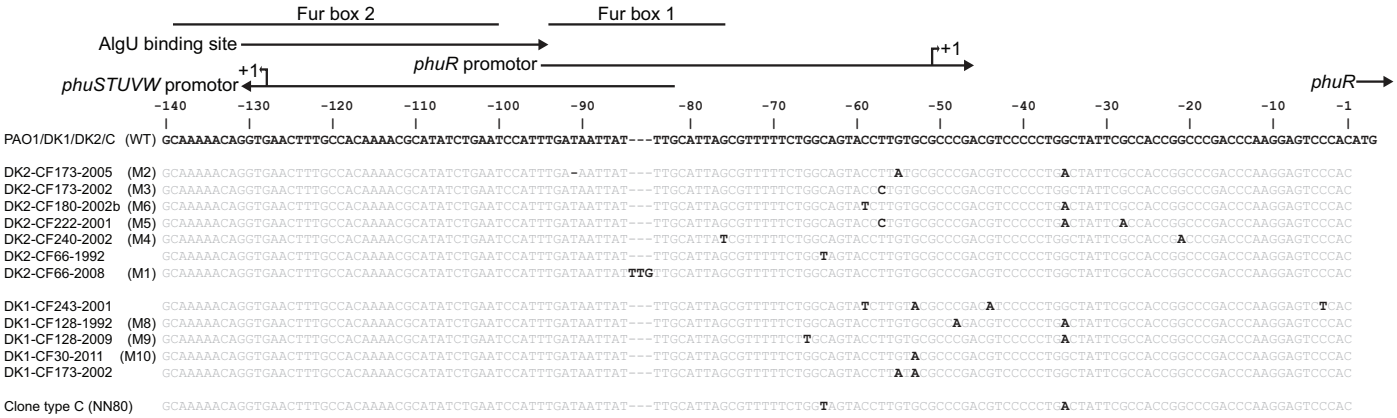


Figure 1: Maximum-parsimonious phylogenetic tree showing the genetic relationship of the 11 DK2 clones included in this study. The phylogenetic tree is adopted from Marvig *et al.* (2013) and is based on mutations identified from whole-genome sequencing. Lengths of branches are proportional to the number of mutations, except in the case of the truncated branch leading to isolate DK2-CF222-2001. Instead, the large number of mutations that have accumulated in isolate DK2-CF222-2001, due to hypermutation caused by a deficient DNA mismatch repair system, is indicated at the end of the truncated branch [2]. Blue and red circles denote mutations in *pvdS* and in the intergenic region upstream of *phuR*, respectively.

All of the 13 mutations are located within a narrow region from position -91 to -21 relative to the start codon of *phuR*, and eight of the mutations are within the annotated promoter region of *phuR* (Figure 2). Furthermore, two positions (positions -35 and -57) were subject to convergent evolution as they were independently mutated in parallel evolving DK2 sub-lineages.



135

136 **Figure 2: Overview of the intergenic region upstream of *phuR*.** The alignment shows homologue
137 sequences from different isolates with genetic variants highlighted in bold. Wild type sequences of *P.*
138 *aeruginosa* strains PAO1, DK1, DK2, and C are shown at the top of the alignment. Abbreviations of
139 sequence alleles from different isolates are indicated in parenthesis (WT and M1-M10). Positions of
140 promoters, Fur boxes, and a binding site of AlgT regulatory protein are indicated with black lines
141 above the alignment. Positions are relative to the start codon of *phuR*.

142

143 *Effect of intergenic mutations on phuR promoter activity*

144 We hypothesized that the mutations, due to their location immediate upstream of
145 *phuR*, could cause an effect on the activity of the *phuR* promoter. To investigate this
146 hypothesis, we cloned the *phuR* promoter region from six of the mutated DK2 clones
147 in front of a luciferase reporter (*luxCDABE*) and chromosomally integrated the
148 transcriptional fusion in *P. aeruginosa* PAO1 at the *attB* site by use of the mini-CTX2
149 derived plasmid pHK-CTX-lux. The transcriptional fusions enabled us to compare the
150 *phuR::lux* expression from the mutated promoter regions (M1-M6) relative to the
151 expression from a construct with a wild type promoter region (WT). A construct
152 without an inserted promoter region was used to correct for background expression
153 from the *lux* gene cassette integration.
154 Measurements of the *phuR::lux* expression at exponential growth (OD_{600}=0.15) in
155 Luria-Bertani (LB) medium revealed that all six mutant alleles (M1-M6) caused a

significant increase in promoter activity with changes in expression from 5 to112 fold (Table 1). The highest increases in expressions (93 and 112 fold) were observed for alleles M1 and M2 originating from clones DK2-CF66-2008 and DK2-CF173-2005, respectively. The M1 and M2 alleles contain a 3-bp insertion and a 1-bp deletion, respectively, in the repressor-binding site (Fur box 1) of the ferric uptake regulator (Fur) known to control the expression of the *phuR* promoter [10]. As Fur mediates a strong repression of *phuR* under iron-rich conditions [10], we find it likely that the indels in the M1 and M2 derived *phuR* promoters alleviates Fur repression.

Origin of *phuR* promoter	Allele	Mean luminescence	SD	Fold change	*P*-value
PAO1	WT	365	±1018	1	
DK2-CF66-2008	M1	34111	±3379	93	0.00021
DK2-CF173-2005	M2	40726	±3422	112	0.00004
DK2-CF173-2002	M3	1879	±3422	5	0.16
DK2-CF240-2002	M4	7584	±496	21	0.00038
DK2-CF222-2001	M5	8968	±610	25	0.00023
DK2-CF180-2002	M6	6723	±701	18	0.00088
DK1-CF128-1992	M8	13329	±1482	37	0.00024
DK1-CF128-2009	M9	12205	±603	33	0.00007
DK1-CF30-2011	M10	9563	±1586	26	0.0011

Table 1: Activities of the *phuR* promoters originating from different clinical isolates of *P. aeruginosa*. Luminescence production from laboratory reference strain PAO1 [17] with *phuR::lux* reporter fusions were measured at exponential growth (OD_{600}=0.15) in Luria-Bertani (LB) medium and normalized for differences in cell density. Mean luminescence production and standard deviations (SD) were calculated from three biological replicates. Statistical analysis concerning the difference between two means was done using a Student's *t* test, and the *P*-values denote the probability of the mutated alleles to have expression equal to wild type (WT).

phuR promoter mutations confer a growth advantage in presence of hemoglobin

The increased expression from the mutated *phu* promoters suggested that the mutations had become fixated due to a positive selection in the CF airways towards iron acquisition from hemoglobin. To test this hypothesis, we replaced the wild type *phuR* promoter of isolate DK2-CF30-1979 with the mutated *phuR* promoter of isolate DK2-CF173-2005 by allelic replacement, and tested whether the constructed mutant strain DK2-CF30-1979-M2 had a growth advantage relative to the isogenic wild type strain DK2-CF30-1979. We chose to test the consequence of the *phuR* promoter mutations in the genetic background of isolate DK2-CF30-1979 as this isolate is an immediate ancestor of isolate DK2-CF173-2005 [18]. For the growth experiment we used a minimal medium supplemented with hemoglobin and apo-transferrin to 'mimic' the conditions of the CF airways.

	Doubling time (hours)		
Growth medium	DK2-CF30-1979	DK2-CF30-1979-M2	***P*-value**
LB	1.27±0.05	1.35±0.07	0.16
ABTGC + 10 uM Fe^{3+}	2.74±0.02	2.69±0.03	0.23
ABTGC + 10 uM Fe^{3+} + 100 µg/mL apo-TF	3.08±0.10	3.07±0.04	0.91
ABTGC + 2.5 uM Hb + 100 µg/mL apo-TF	2.76±0.24	2.13±0.09	0.01

Table 2: Growth rates of strains DK2-CF30-1979 and DK2-CF30-1979-M2 at exponential growth phase in different media. Abbreviations Hb and apo-TF are used for hemoglobin and apo-transferrin, respectively. Note that the ABTGC minimal medium standard recipe was modified so that no iron source other than the one stated in the table was added to the growth medium. Mean doubling times were calculated from three biological replicates. Statistical analysis concerning difference between two means was done using a Student's *t* test, and the *P*-values denote the probability of the two strains to have equal means.

Confirming our hypothesis, we found that the allelic-replacement mutant DK2-CF30-1979-M2 grew significantly faster than its isogenic WT counterpart when hemoglobin

198 was present of as sole iron source (Table 2), while no difference were observed for
199 rich medium and media supplemented with Fe^{3+} as sole iron source.
200
201 *Adaptation towards heme utilization is a general adaptive mechanism*
202 Our results demonstrate parallel adaptation of the DK2 lineage towards hemoglobin
203 utilization in five different CF patients. This indicates that similar selective conditions
204 for heme utilization exist across different patients. Now, we speculated if the fixation
205 of *phuR* promoter mutations is an adaptive mechanism specific to the DK2 lineage, or
206 if *phuR* promoter mutations constitute a general genetic mechanism of *P. aeruginosa*
207 towards heme utilization in the CF airways. To further investigate the generality, we
208 inspected the genomes of two other lineages of *P. aeruginosa* isolated from CF
209 airways.
210 In addition to the DK2 lineage, our previous investigations have revealed another
211 distinct clone type, known as the DK1 clone type, which have also spread among
212 Danish CF patients [19]. We sequenced and analyzed the *phuR* promoter region of
213 five DK1 isolates sampled in years 1992-2011 in addition to an ancestral DK1 isolate
214 from 1973. Whereas the sequence of the *phuR* promoter of the ancestral 1973-isolate
215 (DK1-CF33-1973) was identical to the wild type sequence of strains PAO1 and DK2,
216 all five evolved DK1 isolates had accumulated 1-4 SNPs in the promoter region, and
217 three of the DK1 SNPs were identical to SNPs found in the evolved DK2 isolates. We
218 tested the activity of three of the mutated promoters from the DK1 isolates (M8-
219 M10), and found that all three mutated alleles gave rise to increased levels of
220 transcription similar to what have been observed for mutated DK2 alleles (Table 1).
221 Our result provides strong evidence for convergent adaptive evolution of different
222 lineages of *P. aeruginosa* towards iron acquisition from hemoglobin.

To rule out that the adaptive trait was specific for *P. aeruginosa* CF infections at the Copenhagen CF Center, we analyzed the public available data for the genomic evolution of the *P. aeruginosa* C lineage within patient NN attending the CF clinic at Hannover Medical School, Germany [6]. Interestingly, the C lineage, which has colonized patient NN over at period of over more than 20 years, also accumulated two SNPs in the *phuR* promoter region (Figure 2). Remarkably, the two SNPs are identical to SNPs found in the DK1 and DK2 lineages, and this observation suggests that these mutations were also positively selected for in the host environment.

Selection against pyoverdin secretion might lead to a shift in iron source

The siderophore pyoverdine has previously been found in sputum of CF patients, thus pyoverdine-mediated uptake of iron has been considered important for the survival of *P. aeruginosa* in the CF airways [11]. Nonetheless, we observed both lineages DK1, DK2, and C to have accumulated nonsynonymous mutations in the alternative sigma factor PvdS which is required for pyoverdine synthesis (Figure 1 and Figure 3), and the evolved C clone NN80 was accordingly observed to have lost its ability to produce pyoverdine in contrast to is predecessors (C clones NN2 and NN11) [6].

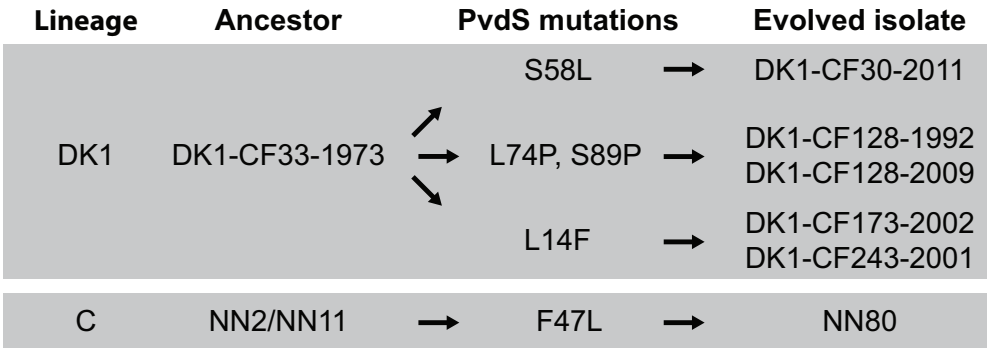


Figure 3: Overview of *pvdS* mutations in the DK1 and C lineages. Mutations that have accumulated in evolved isolates relative to their ancestor are shown. Refer to Figure 1 for overview of *pvdS* mutations in the DK2 lineage.

This led us to examine the production of pyoverdine in the DK1 and DK2 isolates, and we again observed a negative correlation between pyoverdine production and mutations in PvdS (Figure 4). Accordingly, only the ancestral DK1 and DK2 isolates carrying wild type alleles of *pvdS* were able to produce pyoverdine, whereas all isolates carrying mutated alleles of *pvdS* were unable to produce pyoverdine (DK1-CF173-2002 was not tested).

Siderophores are generally regarded as highly immunogenic [20], and selection against pyoverdine production might have driven the accumulation of *pvdS* mutations leading to loss of pyoverdine production in the evolved isolates. At the same time, we observed a positive selection for *phuR* promoter mutations in the CF airways presumable leading to a growth advantage when acquiring iron from hemoglobin. We therefore propose a model in which the CF airways impose a selective pressure on the invading bacteria forcing them to genetically adapt towards a shift to hemoglobin as an alternative iron source. This is of particular interest because inflammation may cause micro-bleeds, which leads to the presence of hemoglobin at the delicate CF lung epithelia in the presence of both host and bacterial proteases [21].

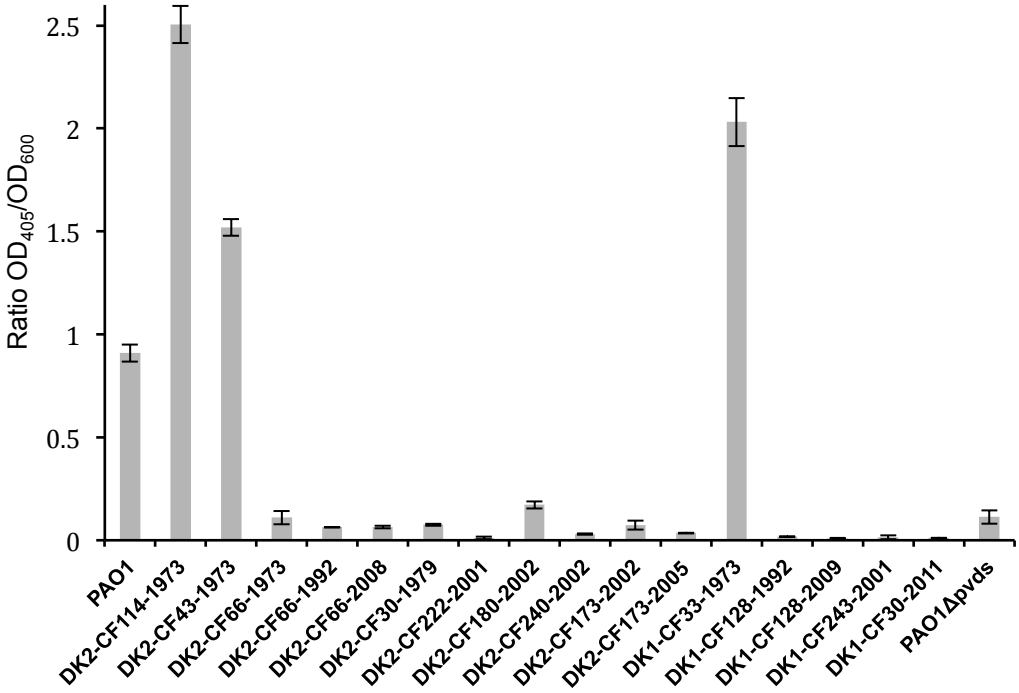


Figure 4: Pyoverdine production in isolates of *P. aeruginosa*. The presence of pyoverdine secreted to the supernatant of bacterial cultures grown in pyoverdine inducing medium was quantified by measurement of absorbance at OD_{405} and normalized against the cell density (OD_{600}). The means and standard deviations calculated from three biological replicates are shown in the barplot.

Conclusions and implications

Our results provide evidence that *P. aeruginosa* fixates *phuR* promoter mutations to genetically adapt toward acquisition of iron from hemoglobin during chronic CF infections. This adaptive trait may be directly selected for due to an abundance of heme-bound iron in the CF lung. Furthermore, we also observed that *phuR* promoter mutations coincided with the loss of pyoverdine production, suggesting that selection for increased heme utilization may be secondary to the loss of the pyoverdine iron uptake system. Therefore, targeting heme utilization might be a promising strategy for the treatment of CF infections.

CF patients commonly experiences iron deficiency, and *P. aeruginosa* possible contributes to iron deficiency by depletion of the host iron storage and causing inflammation [22, 23]. In this regard, expanding of our knowledge of adaptation of *P. aeruginosa* to the CF lung may help to lessen the impact of *P. aeruginosa* infection and improve patient condition.

Materials and Methods:

Bacterials strains and media

Isolates of the *P. aeruginosa* DK1 and DK2 clone types were sampled from Danish CF patients attending the Copenhagen Cystic Fibrosis Clinic. Isolation and identification of *P. aeruginosa* from sputum was done as previously described [24]. Luria-Bertani (LB) broth was used for routine preparations of bacterial cultures. ABTGC minimal medium were composed of 2 g/L $(NH_4)_2SO_4$, 6 g/L Na_2HPO_4, 3 g/L KH_2PO_4, 3 g/L NaCl, 1 mM $MgCl_2$, 0.1 mM $CaCl_2$, 0.01 mM $FeCl_3$, 2.5 mg/l thiamine supplemented with 1% glucose, 0.5% casamino acids. No $FeCl_3$ was added to ABTGC minimal medium for the growth rate experiments presented in Table 2. Pyoverdine inducing medium was composed of ABTGC minimal medium with 50 µM iron chelator 2,2'-dipyridyl (DIPY). *Escherichia coli* strain CC118(λ*pir)* was used for maintenance of recombinant plasmids [25] in medium supplemented with 8 µg/mL of tetracycline. Allelic replacement constructs were transferred to *P. aeruginosa* by triparental mating using helper strain *E. coli* HB101/pRK600 [26]. For marker selection in *P. aeruginosa* 50 µg/mL of tetracycline was used. Genetic techniques were performed using standard methods, and Sanger sequencing was used for verification of genetic construct and allelic replacement mutants.

Sequencing of phuR promoter region and pvdS gene in DK1 isolates

306 Sequencing of DK1 isolates was performed as earlier described [18]. Accordingly,
307 genomic DNA was purified from *P. aeruginosa* isolates using Wizard Genomic DNA
308 purification kit (Promega, Madison, WI-US) and sequenced on Illumina's GAIIx or
309 Hiseq2000 platforms. Reads were mapped against the reference genome sequence
310 using Novoalign (Novocraft Technologies, Selangor, Malaysia) [27], and pileups of
311 read alignments were produced by SAMtools release 0.1.7 [28].
312 *Constrcution of reporter fusions and luminescence measurements*
313 The *lux* gene cassette (*luxCDABE*) was subcloned from the plasmid pUC18-mini-
314 Tn7T-Gm-*lux* [29] fragment into mini-CTX2 [30] using the restriction sites *XhoI* and
315 *PstI* to produce pHK-CTX2-*lux* used for the transcriptional fusion experiments. A
316 220-bp fragment containing the intergenic region upstream of *phuR* was amplified
317 from genomic DNA using Phusion polymerase (Thermo Scientific) with the primers
318 PhuR_F-PstI (5'-GAGACTGCAGAGGCTGGGAGTGCTGCTCAT-3') and
319 PhuR_R-XhoI (5'-ACATCTCGAGAAGGGCGGGGAGAGCGGCAT-3') and
320 ligated with T4 DNA ligase into pHK-CTX2-*lux* after double-digestion of PCR
321 fragment and vector with restriction enzymes XhoI and PstI. The resulting plasmids
322 were introduced into *P. aeruginosa* strain PAO1 by transformation as previously
323 described [29].
324 *Allelic replacement of phuR promoter region in DK2-CF30-1979*
325 A 1296-bp fragment containing the intergenic region upstream of *phuR* was amplified
326 from genomic DNA of DK2-CF173-2005 using Phusion polymerase (Thermo
327 Scientific) with the primers PhuSi_F-XbaI (5'-
328 ACATTCTAGACGGACGTCGCTGGCCTCG-'3) and PhuRi_R-SacI (5'-
329 GAGAGAGCTC-TCTCGTGGCCCTGGCGGTAG-3'). The PCP fragment was
330 ligated into vector pNJ1 [31] after digested with the restriction enzyme XbaI and

SacI. Allelic replacement construct was transferred into strain DK2-CF30-1979 by triparental mating and merodiploid mutants were selected by plating the conjugation mixture on LB agar plates with tetracycline. Colonies were restreaked on selective plates before being streaked on 8 % (wt/vol) sucrose-LB plates without NaCl. Sucrose-resistant and tetracycline-sensitive colonies were restreaked on sucrose-LB plates and screened for the presence of mutated allele by PCR followed by restriction fragment length polymorphism (RFLP) analysis. Positive mutants was finally sequenced by Sanger sequencing at LGC genomics (Germany).

Measurement of growth and luminescence in reporter fusion strains

Overnight culture of the reporter fusion strains were diluted 40 times in fresh LB, and aliquots of 100 µL were transferred to a black (clear bottom) 96 well microtiter plate (Nunc). Three technical replicates were used for each strain, and measurements of growth (OD_{600}) and luminescence were recorded in a Synergy Hybrid H1 reader (BioTek) with six minutes intervals for ten hours and under shaking conditions (200 rpm) at 37°C. Data was analyzed using a custom made script in R version 2.15.2 [32]. Experiment was repeated three times to obtain biological replicates.

Growth rate measurements

Growth rate experiments were carried out in 10 mL Falcon tubes under shaking (200 rpm) at 37°C. Cultures were then inoculated to a starting OD_{600} of 0.005 in 50 mL minimal media in 250 mL baffled shake flasks and measurements of OD_{600} was started nine hours after the inoculation and recorded every 30 min. In the experiment were cells were cultivated in LB, the measurements were started after two hours. The experiment was stopped when the cells reached stationary growth phase typically after around 23 h of growth in minimal media. Growth experiments were repeated three times for each strain in each condition to obtain biological replicates.

Pyoverdine quantification assay

Pyoverdine concentrations were quantified as previously described [33]. All strains were grown in pyoverdin inducing medium for up to OD_{600}>1.5. Cultures were moved into 2 mL micro-centrifuge tubes and centrifuged at 16,000 *g* for two minutes. The supernatants were diluted in 100 mM Tris-HCl buffer (pH 8), and pyoverdine concentrations were quantified by measurement of absorbance at OD_{405}. Finally, the values of absorbance at OD_{405} were normalized against the cell densities (OD_{600}) for each strain. The procedure was repeated for three independent biological replicates.

References:

1. Skaar, E.P., *The battle for iron between bacterial pathogens and their vertebrate hosts.* PLoS Pathog, 2010. **6**(8): p. e1000949.
2. Marvig, R.L., Johansen, H.K., Molin, S., Jelsbak, L., *Genome Analysis of a Transmissible Lineage of Pseudomonas aeruginosa Reveals Pathoadaptive Mutations and Distinct Evolutionary Paths of Hypermutators.* PLoS Genet, 2013. **9**(9): p. e1003741.
3. Rau, M.H., et al., *Deletion and acquisition of genomic content during early stage adaptation of Pseudomonas aeruginosa to a human host environment.* Environ Microbiol, 2012. **14**(8): p. 2200-11.
4. Yang, L., et al., *Evolutionary dynamics of bacteria in a human host environment.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(18): p. 7481-7486.
5. Smith, E.E., et al., *Genetic adaptation by Pseudomonas aeruginosa to the airways of cystic fibrosis patients.* Proc Natl Acad Sci U S A, 2006. **103**(22): p. 8487-92.
6. Cramer, N., et al., *Microevolution of the major common Pseudomonas aeruginosa clones C and PA14 in cystic fibrosis lungs.* Environ Microbiol, 2011. **13**(7): p. 1690-704.
7. Lamont, I.L., A.F. Konings, and D.W. Reid, *Iron acquisition by Pseudomonas aeruginosa in the lungs of patients with cystic fibrosis.* Biometals, 2009. **22**(1): p. 53-60.
8. Poole, K., et al., *Cloning and nucleotide sequence analysis of the ferripyoverdine receptor gene fpvA of Pseudomonas aeruginosa.* J Bacteriol, 1993. **175**(15): p. 4597-604.
9. de Chial, M., et al., *Identification of type II and type III pyoverdine receptors from Pseudomonas aeruginosa.* Microbiology, 2003. **149**(Pt 4): p. 821-31.
10. Ochsner, U.A., Z. Johnson, and M.L. Vasil, *Genetics and regulation of two distinct haem-uptake systems, phu and has, in Pseudomonas aeruginosa.* Microbiology, 2000. **146 (Pt 1)**: p. 185-98.

11. Haas, B., et al., *Siderophore presence in sputa of cystic fibrosis patients.* Infect Immun, 1991. **59**(11): p. 3997-4000.
12. Martin, L.W., et al., *Pseudomonas siderophores in the sputum of patients with cystic fibrosis.* Biometals, 2011. **24**(6): p. 1059-67.
13. Hunter, R.C., et al., *Ferrous iron is a significant component of bioavailable iron in cystic fibrosis airways.* MBio, 2013. **4**(4).
14. Konings, A.F., et al., *Pseudomonas aeruginosa uses multiple pathways to acquire iron during chronic infection in cystic fibrosis lungs.* Infect Immun, 2013. **81**(8): p. 2697-704.
15. Marvig, R.L., et al., *Mutations in 23S rRNA confer resistance against azithromycin in Pseudomonas aeruginosa.* Antimicrob Agents Chemother, 2012.
16. Lieberman, T.D., et al., *Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes.* Nat Genet, 2011. **43**(12): p. 1275-80.
17. Holloway, B.W., V. Krishnapillai, and A.F. Morgan, *Chromosomal genetics of Pseudomonas.* Microbiol Rev, 1979. **43**(1): p. 73-102.
18. Yang, L., et al., *Evolutionary dynamics of bacteria in a human host environment.* Proc Natl Acad Sci U S A, 2011. **108**(18): p. 7481-6.
19. Jelsbak, L., et al., *Molecular epidemiology and dynamics of Pseudomonas aeruginosa populations in lungs of cystic fibrosis patients.* Infect Immun, 2007. **75**(5): p. 2214-24.
20. Wandersman, C. and P. Delepelaire, *Bacterial iron sources: from siderophores to hemophores.* Annu Rev Microbiol, 2004. **58**: p. 611-47.
21. Cosgrove, S., et al., *Pulmonary proteases in the cystic fibrosis lung induce interleukin 8 expression from bronchial epithelial cells via a heme/meprin/epidermal growth factor receptor/Toll-like receptor pathway.* J Biol Chem, 2011. **286**(9): p. 7692-704.
22. Pond, M.N., A.M. Morton, and S.P. Conway, *Functional iron deficiency in adults with cystic fibrosis.* Respir Med, 1996. **90**(7): p. 409-13.
23. Reid, D.W., et al., *Iron deficiency in cystic fibrosis: relationship to lung disease severity and chronic Pseudomonas aeruginosa infection.* Chest, 2002. **121**(1): p. 48-54.
24. Hoiby, N. and B. Frederiksen, *Microbiology of cystic fibrosis*, in *Cystic fibrosis*, M. Hodson and D. Geddes, Editors. 2000, Arnold: London, United Kingdom. p. 83-107.
25. Herrero, M., V. de Lorenzo, and K.N. Timmis, *Transposon vectors containing non-antibiotic resistance selection markers for cloning and stable chromosomal insertion of foreign genes in gram-negative bacteria.* J Bacteriol, 1990. **172**(11): p. 6557-67.
26. Kessler, B., V. de Lorenzo, and K.N. Timmis, *A general system to integrate lacZ fusions into the chromosomes of gram-negative eubacteria: regulation of the Pm promoter of the TOL plasmid studied with all controlling elements in monocopy.* Mol Gen Genet, 1992. **233**(1-2): p. 293-301.
27. Krawitz, P., et al., *Microindel detection in short-read sequence data.* Bioinformatics, 2010. **26**(6): p. 722-9.
28. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
29. Choi, K.H. and H.P. Schweizer, *mini-Tn7 insertion in bacteria with secondary, non-glmS-linked attTn7 sites: example Proteus mirabilis HI4320.* Nat Protoc, 2006. **1**(1): p. 170-8.

30. Hoang, T.T., et al., *Integration-proficient plasmids for Pseudomonas aeruginosa: site-specific integration and use for engineering of reporter and expression strains.* Plasmid, 2000. **43**(1): p. 59-72.
31. Yang, L., et al., *Polysaccharides serve as scaffold of biofilms formed by mucoid Pseudomonas aeruginosa.* FEMS Immunol Med Microbiol, 2012. **65**(2): p. 366-76.
32. Team, R.D.C., *R: A Language and Environment for Statistical Computing.* http://www.R-project.org, 2009.
33. Imperi, F., F. Tiburzi, and P. Visca, *Molecular basis of pyoverdine siderophore recycling in Pseudomonas aeruginosa.* Proc Natl Acad Sci U S A, 2009. **106**(48): p. 20440-5.

Infection Microbiology Group
Department of Systems Biology
Technical University of Denmark

Matematiktorvet
Bygning 301
2800 Kgs. Lyngby
Tel. 45 25 26 00
Fax 45 88 49 22
www.bio.dtu.dk

ISBN: 978-87-91494-58-1